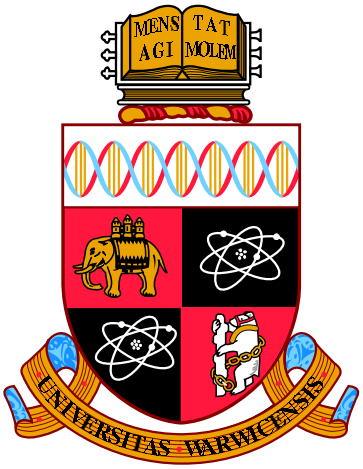

Facial Expression Recognition under Harsh Lighting using High Dynamic Range Imaging

by

Emmanuel Omotayo Ige (B.Sc Maths, M.Sc Computer Science)

Thesis

Submitted to the University of Warwick for the degree of

Doctor of Philosophy

Warwick Manufacturing Group

October 2016

THE UNIVERSITY OF WARWICK

Contents

Acknowledgments

First, I thank God Almighty, the source of my inspiration and success throughout the PhD program.

Many thanks to my supervisors Professor Alan Chalmers and Dr Kurt Debattista for taking me on in the darkest time, giving me the opportunity to continue my PhD. I am forever grateful for the invaluable advice, encouragement and support. God bless!

For the good memories and time spent in the Visualisation group - Dr Carlo, Dr. Tom, Dr. Ali, Ratnajit, Stratou, Josh, Jon, Demalyah, Tim, Martin, Pina, Rusella, my thanks to you all.

To other friends around during my stay - Dr. Nentawe, Chinedu, Daniel, Ade, Martins, Fatima, Alaa, Elaine, you guys are jolly good friends.

Special thanks to my wife, Francisca and children, Laura, Lewis and Lionel for their love and sacrifices. Thanks for being patient when I am always away in the lab.

My sincere gratitude to Bolanle Obash for her encouragement and financial support when most needed.

My family have always been a source of support and prayers to me - My mum, brothers and sisters, I remain grateful.

Finally, this thesis is dedicated to my dad Julius Ige of blessed memory. You would have been happier to read my thesis and witness my graduation. May your soul rest in peace. Amen!

Declarations

The work in this thesis is original and no portion of work referred to here has been submitted in support of an application for another degree or qualification of this or any other university or institute of learning.

Signed:

Emmanuel Ige

List of Tables

List of Figures

Abbreviations

HDR - high dynamic range

LDR - low dynamic range

Opt-exp - optimal exposure

FER - facial expression recognition

FR - face recognition

SURF - speeded-up robust features

BOF - bag of features

LBP - local binary pattern

SVM - support vector machine

CNN - convolutional neural network

FACS - facial action coding system

AU - action unit

AHE - adaptive histogram equalisation

NDCT - normalised discrete cosine transform

WF - wiener filter

GC - gamma correction

JAFFE - Japanese female facial expression

mJAFFE - modified Japanese female facial expression

SFEW - static facial expression in the wild

Log - logarithmic

Ovh - overhead light

LL - left light

BL - back light

TMO - tone mapping operator

DA - display adaptive

Dr - drago

Re - reinhard

SBS - surf, bag of feature, svm

PCA - principal component analysis

ICA - independent component analysis

LDA - linear discriminant analysis

CR - contrast ratio

EV - exposure value

HVS - human visual system

Publications

The results of the following publications have been summarised within this thesis:

Peer Review Conference Papers

1. **Ige, E. O.**, Debattista, K. and Chalmers, A., 2016, June, Towards High Dynamic Range (HDR) Based Facial Expression Recognition under Complex Lighting, in proceedings of the 33rd International Conference on Computer Graphics, (CGI'16), ACM, Heralklion, Greece, June 28 - July 01, pp. 49-52.

2. **Ige, E. O.**, Debattista, K., Mukhajee, R. and Chalmers, A., 2016, Exploring Face Recognition under Complex Lighting Conditions with High Dynamic Range (HDR) Imaging, in proceedings of the International Conference on Computer Graphics and Visual Computing (CGVC'16), September 2016.

Journal Paper

Ige, E. O., Debattista, K. and Chalmers, A. (Submitted) 2017, Evaluating Facial Expression Recognition under Harsh Lighting with an Enhanced High Dynamic Range (HDR) Database, in Journal of Visual Communication and Image Representation.

Other Conference Papers

1. **Ige, E. O.**, HDR Based Facial Expression Recognition Under Complex Lighting, WMG Doctoral Research and Innovation Conference, IMC and IDL, University of Warwick, 21 June 2016.

2. **Ige, E. O.**, Comparison of Subjective Methods for Image Quality Assessment, WMG Doctoral Research and Innovation Conference, IMC and IDL, University of Warwick, 10-11 July 2014.

3. **Ige, E. O.**, Facial Expression Recognition: Is the Technology Threat to Privacy? Presentation at Cafe Workshop Scientific Doctoral Presentation Series, 12th December, 2012, postgraduate research skills of the University of Warwick.

Poster Presentations

1. The Science and Psychology of Facial Expression Recognition of Emotions, presented at the International School on Biometrics for Secure Authentication: Understanding Man-Machine Interactions in Forensics and Security Applications, Algehro Italy, June 2012.

2. Does Facial Lighting Affect Interpretation of Facial Expression of Emotions? 2nd Workshop on Visual Image Interpretation in Humans and Machines (ViiHM), Bailbrook House, Bath, Somerset, 1-2 July 2015.

3. Can High Dynamic Range (HDR) Images be Adopted for Facial Expression Recognition of Emotions? 3rd Workshop on Visual Image Interpretation in Humans and Machines (ViiHM), Bailbrook House, Bath, Somerset, 12âĂŞ13 July 2016.

Abstract

Facial information can reveal the emotional status of individuals. Although traditional cameras can capture this information, such cameras struggle to acquire the necessary information in extreme lighting conditions. This thesis aim to investigate whether High Dynamic Range (HDR) imaging can capture human facial expression under complex lighting conditions, and in doing so, enhance Facial Expression Recognition (FER) performance. The techniques presented in this thesis focus on developing a baseline for images captured in scenes with harsh lighting conditions, where Low Dynamic Range (LDR) images have difficulty capturing the full range of light in a single exposure. The thesis considers unprocessed images and a variety of pre-processing methods to examine whether reducing the impact of large lighting variations could improve the quality of an input image.

In addition, realistic facial data plays a key role in validating facial expression analysis systems. Today, the majority of FER algorithms are evaluated only on images generated in highly controlled laboratory environments. The variability of a facial appearance in an image could be dominated by changes in head pose and illumination conditions. This can effectively hide features that are necessary to discriminate different subjects or different facial articulations. New HDR imaging techniques are thus introduced to help ensure that all the details in a scene is captured no matter what the lighting conditions present, and all this detail is then available to the FER algorithms. This is also investigated on Face recognition algorithms.

Chapter 1

Introduction

"Most of the problems in life are because of two reason - we act without thinking or we keep thinking without acting."

In a world of over six billion different faces, each human face is unique. Each face is a source of rich information. It can reveal character, cultural identity, age, gender, beauty, personality, emotions. Figure 1.1 shows sample human faces. The face can convey different information, it can sometimes be used as a source of meaningful information and can also be a confusing source of information. For example, a prematurely wrinkled face in a young person could mislead interpretation of expressions of emotions. Furthermore, the face is a location for sensory input and communication output [EFE13]. Charles Darwin [Dar72] described a face as an important source of evidence to describe human character, for example, the facial muscles activities aid the making of expressions. Through the facial muscles, humans are able to show and share social information with others and even engage in non-verbal communication. This can provide rich information about people's intention or mind set [SB10].

Although most of the facial information is passed on a subconscious level, we still rely on the interpretation of facial expressions to determine emotional state in order to form a prediction of the reaction. A number of questions about

Figure 1.1: Sample human emotional faces

emotion perception exist, for example, can people be judged based on their facial expressions? Does a smile indicate a person is happy or insincere? Does squinting show concentration or mistrust? Facial expression is displayed when muscles beneath the facial skin move. This movement can reveal the current focus of attention, synchronise a dialogue, and signal comprehension or disagreement. Furthermore, it can convey social and emotional information between humans and, as some researchers believe, these are the primary means of non-verbal communications. According to [Meh08], in a conversation, whether the listener feels liked or disliked amounts only to 7% of the spoken word, 38% on vocal utterances and 55% on facial expressions based on emotions.

Humans perceive and interpret facial expressions almost seamlessly by way of observation and evaluation of subtle changes in the key facial features, such as the eyes, eyebrows, nostrils and lips. This has been widely studied from the perspective of human psychology. These studies have given rise to several theories on how to encode, represent and interpret facial expressions. While emotions

can be extracted and interpreted from facial expressions, its interpretation relies heavily on the information collected from the face. For example, as we laugh or frown we are putting our emotions on display, allowing others to infer our current emotional state. Therefore, the knowledge of how much information the face provides is necessary to interpret emotions. Research reveals that emotions control our actions. It is thought that emotions have developed via the adaptive values in human fundamental life styles [Ekm92]. The human affective state and the associated behavioural expressions constitute an important part of our life. It has been established that they are reflected in the way we behave, make decisions and communicate with others, which means that the affective state of humans influences their actions and of others around them [RWHJ13].

Researchers in the face domain have raised a series of theories to answer the question, *"can a computer recognise a face as the same when presented under different lighting conditions?"* This thesis will carry out a study in order to provide an answer to the question. In doing so, the focus is restricted to one area of facial information - facial expression. The aim is to concentrate upon questions on how the face provides information about facial expressions and what are the challenges associated with the study of facial expressions. For example, *does the face provide accurate information about expressions when presented under harsh lighting conditions*?

This thesis proposes HDR imaging techniques to aid the analysis of facial expression recognition of emotions with images captured under harsh lighting conditions such that the process of image pre-processing can be avoided in the FER processing system.

1.1 Facial Expression Recognition

In order to recognise emotions from facial expression in a presented image/video, a variety of image processing techniques are employed. The choice of these

techniques relies heavily on a number of factors such as, the type of image/video data, condition of the image/video data and the application areas involved. To put that in perspective, first, an understanding of the condition of the data provided is optimal, followed by a robust feature identification and extraction technique for extracting useful features. On top of that, the feature extraction technique should be able to provide adequate features for the learning/classification techniques.

FER is highly beneficial for diverse application areas. With FER, we can test the impact of any content, product or service that is associated with the elicitation of emotional arousal and facial responses. For example, in physical objects such as food or packages, videos and images, sounds, smell, sense of touch stimuli and so on.

Recently, facial expressions of emotion recognition techniques have been adopted in academic research and commercial fields in prominent areas such as media testing and advertisement, consumer neuroscience and neuromarketing, clinical psychology and psychotherapy, medical applications and plastic surgery, software user interface and website design, engineering of artificial social agents (avatars), educational technology and workplace stress assessment. It is this wide range of applications that has produced a surge of interest in machine analysis of facial expressions.

At the moment, even the best algorithms are reliant on a number of processing techniques to perform better than human judgement [Kir09]. However, human judgement can sometimes be unreliable particularly when urgent decisions are needed. Research in the domain of FER uses a variety of tools to assure reliability, for example, when recognising people's emotions in environments where even the best camera struggle to capture enough information under extreme changes in brightness. Based on this, in this thesis, an investigation will be conducted seeking better ways of achieving robustness to issues of image lighting conditions. In addition, face recognition (FR) will also be explored. This is because

FR and FER systems use similar processing methods. For example, face detection is a technique used in both FR and FER systems. Therefore, in the course of this research, an investigation will be carried out into how FR performance can be improved when presented with images under complex lighting conditions.

1.1.1 Facial Expression Recognition Pipeline

One of the contributions of this thesis is the exploration of the existing facial expression recognition techniques used for dealing with images affected with harsh lighting conditions. Therefore, there is the need to introduce the FER pipeline. A generic FER system pipeline involves, data/video capturing, feature

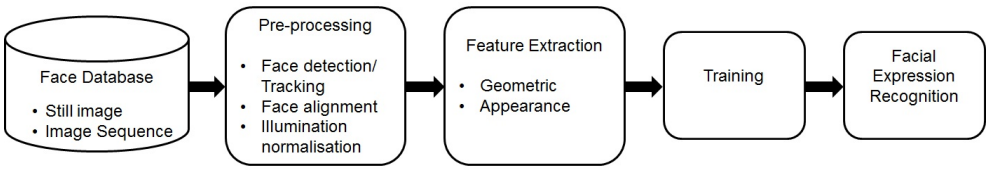

Figure 1.2: Facial expression recognition pipeline representing the stages through which images and videos pass from input to output. Different processing is involved at each stage depending on the intended application.

identification/extraction and learning/classification as shown in Figure 1.2. Each stage of the basic FER pipeline constitutes a set of established functions and operations which takes an image(s), pre-processes it (if needed), performs feature extraction and makes classification based on defined classes depending on the intended applications. Each stage is examined in more detail in the course of this thesis. The focus is on avoiding the need for the pre-processing stage.

1.2 Face Recognition

As mentioned earlier, faces captured can be affected by environmental factors such as harsh lighting, limited field of view, poor image quality, occlusion and even facial expression, not needed for recognition. These factors are difficult to

fully control in real-world environments. Thus, face recognition systems suffer a significant performance drop with these constraints. Amongst these factors, the focus in this thesis is on harsh lighting conditions. In Chapter 5, FR under complex lighting conditions is discussed.

Most FER and FR software has relied on images captured from an existing database. To be effective and accurate, the images captured need to be of a face with little variance of light as in the case of images in the database. In most instances, if the images were not taken in a controlled environment, small changes in light could reduce the effectiveness of the system. For example, Figure 1.3 shows FER/FR processing stages. In order to give more understanding to this, issues of image

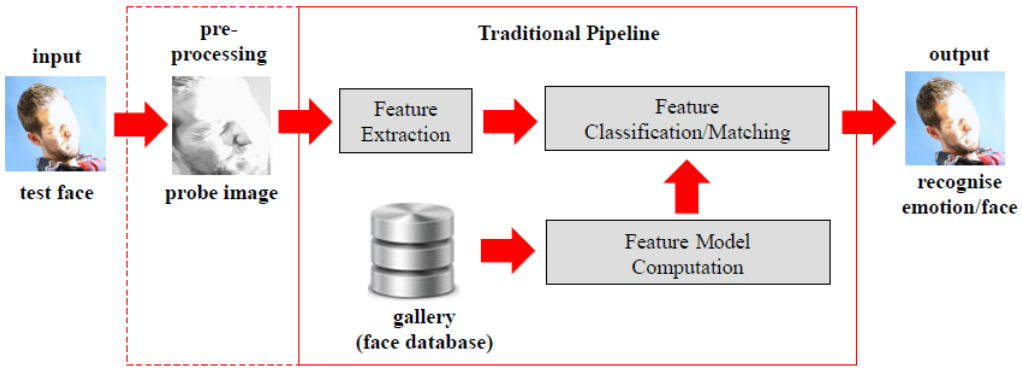

Figure 1.3: Facial expression recognition/Face recognition processing stages.

lighting conditions will be explored further.

1.3 Image Lighting Conditions

Images often contain a variety of effects caused by changes in lighting. Scene light plays an important role when capturing images for the purpose of image processing applications. This is because, diffuse lighting can produce destructive superposition of light and shadow, which can result in a permanent loss of information [WLH*07]. Sample faces under harsh lighting is shown in Figure 1.4. In our case, most images captured for FER purposes are made ideal within

Figure 1.4: Sample faces under harsh lighting

controlled lit scenes, where faces of the subjects are perfectly lit and stable to ensure flawless capture. Certainly, the recovering of the geometry and texture of presented faces is very important to any image processing system [WLH03]. This can particularly be difficult when the images are captured under harsh lighting conditions, and even more challenging when cast shadows, saturated areas and partial occlusions are present. Harsh lighting is produced when dark shadow is formed due to large contrast between the light and dark areas in an image. However, it has been reported that approximation errors become large in the presence of images under harsh lighting conditions [WLH*07]. Adini et al [AMU97] show that changing illumination direction could influence the perception of object characteristics such as 3D shape and location. Therefore, changes in lighting result in large image differences. Such changes can be even larger when varying the identity of the subject.

1.3.1 Image Enhancement

Image enhancement is useful where the subjective quality of an image is important for human or machine interpretation. It is a process of adjusting images to a more suitable form for display or further image analysis. In this thesis, image enhancement, a form of pre-processing is based on image contrast enhancement. Contrast is an important factor in any subjective evaluation of image quality. Contrast is created by the difference in luminance reflected from two adjacent surfaces [SMCD03]. It is thus the difference in visual properties that makes an

image distinguishable from another image and the background. An example of a non-linear contrast enhancement is histogram equalisation in which the original histogram is redistributed to produce a uniform population density of pixels along the horizontal axis. This stretch applies the greatest contrast enhancement to the most populated range or brightness values in the original image. In Figure 1.6, the intensity range of the darkest values are preferentially stretched, which results in maximum contrast. The uniform distribution stretch strongly saturates darkness values at the sparsely populated light and dark regions of the original histogram.

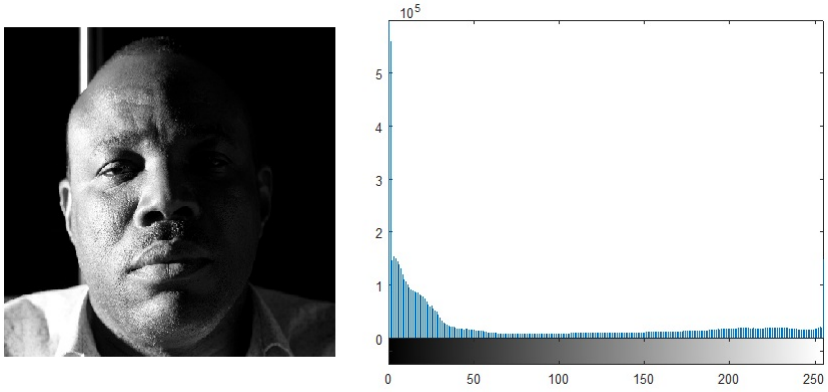

Figure 1.5: Original image with half of the face in shadow. With most values in the histogram towards the dark region of the intensity range

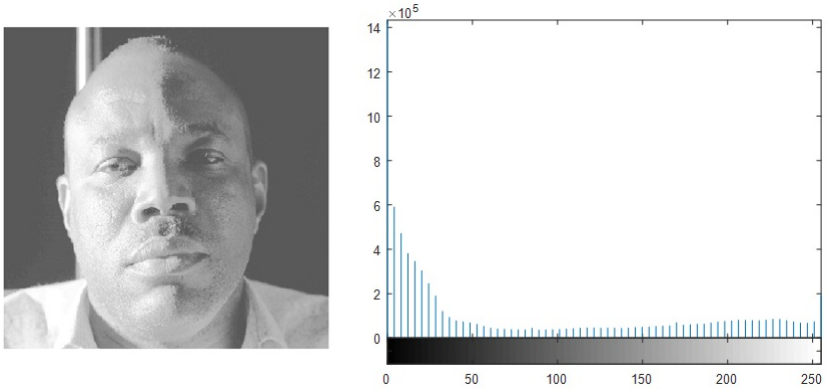

Figure 1.6: Contrast of the original image adjusted using histogram equalisation. The histogram equalisation function tries to match a flat histogram with 64 bins, which is the default behaviour

1.4 High Dynamic Range Imaging

Traditional imaging techniques, known as Low Dynamic Range (LDR) imaging is able to capture a limited dynamic range of a scene. This results in over-exposed (or white) pixels and under-exposed (or black) pixels within the LDR image at places where details would be expected. Where the dark and bright regions of a scene can be recorded at the same time onto an image/video, for example, when capturing images on a sunny day through a window, choices need to be made on whether to capture either the background sky and highlights or capture the scene details in the shadow, although both will be visible to the observer, as shown in Figure 1.7. HDR imaging techniques typically capture multiple exposures of the same scene

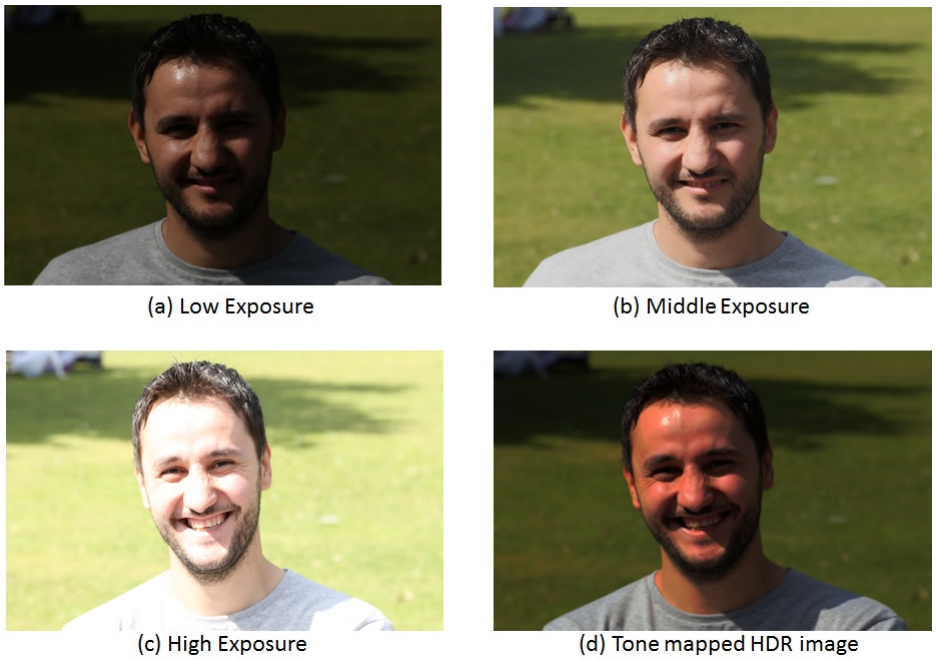

Figure 1.7: (a) In low exposure it is possible to see information in the highlights, while the rest of the detail is lost, such as in the colour of the grass. (b) The middle exposure shows clipping of the image both in the highlights and shadows but it preserves most of the data. (c) High exposure contains all shadow details but the rest of the image is overexposed. (d) The three exposures combined into a single HDR image which preserves all the information visible to the human eye.

and combine these into a single image. This allows all the scene data to be captured; there are no more under or over-exposed pixels. HDR images can be converted to an LDR image through a process known as tone mapping [BADC11].

1.5 Research Challenges

In spite of the fact that in the FER research community a number of different approaches have been developed, the recognition of facial expressions in complex lighting scenarios remains an unresolved problem. There are several instances where due to different environmental factors, it is not always possible to get a suitably illuminated image. In such cases, machine learning and object recognition techniques may lose their effectiveness. In this thesis, the use of HDR to improve FER will be investigated to overcome this problem. The research challenges are:

1. In reality, most of the image data captured will not be under 'ideal' studio lighting conditions, where the face of the subject is perfectly lit and stable to avoid imperfections. However, most FER systems have been developed for environments with homogeneous lighting condition, such as outdoors, indoors, through a window or in a low light visibility area.

2. When images are captured under changing light conditions the differences in lighting conditions contribute more to image differences than changes in facial features and even more to variations in emotions [PSO*07]. Changing lighting conditions can pose huge challenges to computer vision techniques. Thus, using traditional capturing techniques, for example in the 'wild' typically involves a trade-off within the extreme lighting conditions. When a face is illuminated with different light sources, the face may look different, thereby the facial images of the same person can appear non-identical under different lighting conditions [BJ11] as shown in Figure 1.8. Such differences can confuse emotion recognition systems [PSO*07]. We therefore consider

this as one of the essential motivations for this thesis.

Figure 1.8: Sample images of the same person in the same pose under different lighting conditions

3. The effects of lighting in smaller image regions can appear homogeneous [WLH*07] and can become worse if the majority of the pixels in the region are problematic. Therefore, the full image information cannot be correctly recovered, for example, in cast shadows, saturated areas or when there are large lighting estimation errors, Figure 1.8 (left) and Figure 1.9 (right). Thus, arbitrary lighting conditions, occlusion, low image resolution, pose, differences in expressions across different cultures and other parameters, such as camera capability, can make FER tasks even harder [PSO*07].

4. Analysis of facial expressions has been a challenging research area, even with the development of different FER approaches in the last few years. For instance in the real world applications such as visual surveillance, video conferencing or tracking of a person's emotions are some applications that requires a FER system that works adequately on images under challenging lighting conditions. There exist many methods of FER but very few of them provides results or can work in challenging scenarios.

Figure 1.9: Sample facial expression images under harsh lighting. From left: cast shadow, partial occlusion and saturated areas.

1.6 Aims and Objectives

1.6.1 Aims

The research reported in this thesis aims at targeting the deficiencies and limitations of the existing FER systems with images captured under harsh lighting conditions. Most FER studies have directed a great deal of attention on data from images with relatively uniform illumination. The key challenge will be how to achieve optimal performance via pre-processing, feature extraction and classification, under harsh lighting conditions. Due to the growing need to recognise emotions or track how people react through their facial expressions for applications requiring images captured under rapid light changes, the current systems built on lab collected data are not designed to generalise to such scenarios. Therefore, the general objective of this thesis is on improving the performance of facial expression applications in scenes where harsh lighting conditions can be expected and full image information is not necessarily given, and where the traditional LDR imaging techniques struggle to cope with the lighting changes.

1.6.2 Objectives

The following are the specific objectives of the thesis:

- Explore the existing facial expression recognition techniques.

- Investigate how to recognise emotions from images showing facial expressions in harsh lighting conditions without first pre-processing the input image.
- Build a FER system with high level of insensitivity to changes in light. In particular, we consider whether HDR provides a significant improvement for FER systems due to its capability of allowing the capture and manipulation of natural scene luminance ensuring all details in a scene are captured no matter what the lighting conditions.
- Find out whether tone mapping methods are sufficient for improving performance of facial expression recognition systems.
- Validate the level of performance that can be achieved with HDR methods under different lighting conditions for FER tasks
- Examine whether lighting a face from different directions does have a significant effect on recognition.

1.7 Research Questions

The thesis attempts to answer the following research questions:

- How are current methods used for facial expression recognition problems coping in the presence of data presented under harsh lighting conditions?
- Are some of the image enhancement techniques adopted able to boost facial expression recognition tasks? If so, can this performance be enhanced?
- Can a HDR method help to reduce loss of information when used as an alternative to image enhancement?
- Is tone mapping an HDR image to an LDR image sufficient to get the desired performance from the FER?

1.8 Approach

The steps to answer the research questions are as follows:

1. Sets of images will be captured under harsh lighting conditions to investigate whether pre-processing tasks holds value for FER.

2. FER experiments will be conducted using the original images versus the pre-processed images.

3. A facial expression dataset of HDR images will be created. Machine learning methods, Local Binary Pattern (LBP), Speeded-up Robust Features (SURF) and Support Vector Machines (SVMs) will be used to conduct emotion recognition experiments.

4. HDR will be investigated as to whether it can benefit FR systems.

1.9 Thesis Outline

The thesis consists of eight chapters which are organised as follows:

- Chapter 2 provides an overview of the methods and techniques used for facial expression recognition analysis. It covers face recognition, emotion recognition, classification and facial expression databases.

- Chapter 3 details the fundamental concepts of High Dynamic Range imaging. It gives the difference between LDR and HDR imaging. Different types of tone mapping operators (TMOs) are discussed. This is followed by discussion on the elements of HDR imaging as it relates to capture and post processing.

- Chapter 4 discuses the overall methodology and presents the experiments used for creating the HDR facial expression database and for understanding

whether HDR can be useful in addressing the limitations of the LDR imaging methods in handling scenes with harsh lighting conditions.

- Chapter 5 investigates the effects of HDR tone mapping methods in the presence of complex lighting conditions for FER. A database was created with strong shadow affecting half of the face. Image processing methods were used to conduct emotion recognition in order to create a bench mark for further experiments.

- Chapter 6 presents an investigation exploring whether HDR can benefit face recognition. A new database was created under three different lighting scenarios. Existing face recognition methods were used to test HDR imaging performance on face recognition systems and, comparisons were made for further validation of the new HDR facial expressions database.

- Chapter 7 validates the new HDR database by investigating the performance of FER tasks on the database. Two facial expression algorithms were implemented in order for performance comparison. The results of these are used to create a benchmark for the created HDR facial expression database for the research community.

- Chapter 8 presents the overall thesis conclusions. In addition, the thesis contributions to the field of computer vision and image processing are given and areas for further studies presented.

Chapter 2

Facial Expression Recognition System Review

"The secret of change is to focus all of your energy, not on fighting the old, but on building the new." Socrates

2.1 Introduction

Facial expression reveals not only the facial motions, but also subtle changes under lighting and appearance [WLH*07]. These changes are important details for visual cues, but could also be difficult to synthesise.

A number of traditional approaches for processing images under harsh lighting conditions have been proposed: normalisation, illumination extracting or representation, such as spherical harmonics or illumination cone [LHK05, ZS03, SGCZ03, AHP06, PYL08]. Although, the methods generally focus on reducing the effect of lighting, their resilience to difficult lighting variations is still quite limited. We therefore argue that the task of FER could be done in a more conducive manner.

2.2 Facial Expression Analysis

Facial expression analysis can be categorised into video-based and image-based. The focus of this thesis is on the image-based. Facial expression analysis usually starts with face detection in videos or images. However, this may not be the case when using already pre-processed data. Face detection is an essential pre-processing task carried out before performing face/facial expression recognition. Another pre-processing task is face normalisation. Closely linked to face detection is face segmentation. Some applications may require face segmentation as a necessary step, whereas some may not. With FER of emotions, the use of facial parameter coding systems is very important as a metric for measuring facial expressions. To evaluate facial expressions data for emotion recognition, the facial features need to be extracted. This can be carried out using three methods - muscle based, model based and motion based. More than two of these methods can be combined as hybrid methods. The final step of a FER system is facial feature classification into different expressions.

Facial expression analysis can be summarised in three different ways [WLH*07]

- Tracking of facial electromyographic activity (fEMG)
- By live observation and manual coding of facial activity, using the facial action coding system (FACS)
- By automatic facial expression analysis using computer-vision algorithms, which is the focus of this thesis.

The next three sections will be used to describe face detection, face normalisation and segmentation.

2.2.1 Face Detection

Face detection is a procedure for finding the position of a face in an image/video [KMM10]. The traditional (commonly used) algorithm for face detection is the Viola Jones cascade classifier [VJ01], a method that uses the Haar-like features, which are efficiently calculated using the integral image. The results of face detection enables tasks in facial analysis such as facial expression recognition and face recognition on focused image areas. The goal of face detection or tracking module is to determine in a given arbitrary image/video whether or not there are any faces in the image/video and if present, return its location and the extent of each face. Locating a face within the image/video is referred to as face detection or face localisation, while face tracking is the tracking of a face across the different frames of a video sequence [KMM10]. Figure 2.1 shows an example of face detection in an image. .

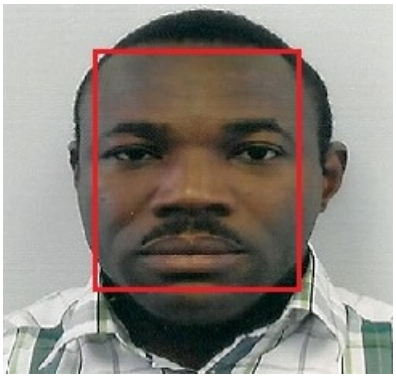

Figure 2.1: Face detection with the Viola-Jones algorithm which uses a cascade object detector to detect the presented face [VJ01].

A face detection algorithm works in a similar way as the technology used in Android smartphones and iPhones cameras. Usually, it results in a box framing the detected face. Although, face detection and tracking processes appear trivial with humans, sometimes data presented to the computer system may be characterised by difficulties such as, unstable scene lighting conditions, background occlusion,

facial pose, orientation, etc [LZY12]. The authors Fanelli et al [FDG*13], proposed a face detection algorithm based on the Haar-like features extracted from a range of data used as features in random forest decision nodes. This face detection algorithm was used to demonstrate the ability of handling large rotations, partial occlusions and the noisy depth data acquired using commercial sensors. Recently [FB15] introduced a novel statistical framework for face detection and facial expression recognition. The framework is based on the Dirichlet process mixture of generalised Dirichlet distributions, which they used to model the local binary pattern features. They adopted a localised feature selection scheme, which they reported showed superior performance over the traditional method. They illustrated their results using synthetic data adopted to carry out face detection and FER [FB15].

2.2.2 Face Normalisation

Face normalisation plays an important role in facial analysis. In the image processing domain, normalization is a process involving changing the range of pixel intensity values [HYH*05]. The purpose of face normalisation is to bring the image into a range that is more familiar or normal to the scene. Normalization can sometimes be referred to as contrast stretching or histogram stretching [GW08]. In digital signal processing (DSP), face normalisation is referred to as dynamic range expansion [GW08]. Contrast stretching is a simple image enhancement technique employed to improve image contrast by stretching its range of intensity values, where a scaling function is usually applied to the image pixel values. With contrast stretching, the upper and lower pixel value limits are usually specified over which the image is to be normalised. Most often, the limits are the allowable minimum and maximum pixel values for the image [HYH*05]. To put this in perspective, let $X_{original} = \{X_{original}[n_1, n_2, n_3] \mid 1 \leq n_1 \leq N_1^{'}, 1 \leq n_2 \leq N_2^{'}, 1 \leq n_3 \leq 3 \mid\}$ denote the original image, and $N_1^{'}$ and $N_2^{'}$ are the original sizes of the presented

image and n_3 represents the three colour channels (RGB) for colour images. For example, an image $X = \{X[n_1, n_2, n_3] \mid 1 \leq n_1 \leq N_1, 1 \leq n_2 \leq N_2, 1 \leq n_3 \leq 3 \mid\}$ can be used to represent a certain facial image containing a face. Conversely, with a greyscale image the colour components are ignored, $X_{original}[n_1, n_2]$. It is worth noting that for an 8 bits grey scale image the lower and upper limits are between 0 and 255. The traditional normalisation method scans the image in order to find the lowest and highest pixel values present in the image. Afterwards, each of the pixel is scaled to a new value as presented in the equation below [BKW*08]:

$$X_{norm}[n_1, n_2] = (X[n_1, n_2] - x_{min})(\frac{255 - 0}{x_{max} - x_{min}}) + x_{min} \tag{2.1}$$

$$x_{max} = max_{\forall n_1, \forall n_2}\{X[n_1, n_2]\} \tag{2.2}$$

and

$$x_{min} = min_{\forall n_1, \forall n_2}\{X[n_1, n_2]\} \tag{2.3}$$

where x_{min} and x_{max} are the lowest and highest pixel values present in the image.

Variability of facial appearance in image data could be influenced by changes in either head pose or lighting conditions as these, have the ability of hiding information useful for discriminating between subjects or the different facial clarity. After estimating the relative position of the features, the image is then warped so that features are approximately located in the predefined positions of the image. In order to reduce computational constraints in the processing task, normalisation is usually performed in the feature space and not in the original image domain. To overcome the effects of head pose, scaling, and in-plane normalisation is carried out to achieve head pose correction through the estimation of salient facial features, such as eyes, nose and mouth. Breitenstein et al proposed a morphable 3D face model parameters [BKW*08] to re-render the face from a new view point, with parameters such as shape, texture, pose and lighting

condition. These are iteratively estimated by minimising the differences in pixel colour between the image and the rendered morphed model. Similarly, changes in facial light has been shown in [AMU97] to introduce bigger changes in the image facial appearance than the head pose. Various techniques have been proposed to address this problem. The authors in [Sha97] observe that images generated under unstable lighting conditions can be represented in the 3D linear subspace.

2.2.3 Face Segmentation

Face segmentation is essential for face classification for facial expression analysis. It involves partitioning in order to separate the facial pixels from all other background pixels [PP06]. The simplest methods of face segmentation are based on the skin colour, although, face segmentation could provide more satisfactory results. However, some of the pixels may be incorrectly labelled due to the issues of changing light, specular reflections or background clutter. Recently, face segmentation algorithms have been designed incorporating, in addition to the data driven part, a face model, such as the information about the variability of face shape in the form of point distribution model [JWYh09]. Pantic and Patras [PP06] present a system for automatic recognition of facial action units. Their algorithm performs both automatic segmentation of an input video into facial expressions pictured and recognition of temporal segments: onset, apex, offset of 27 Action Units (see below) occurring alone or in a combination in the input face profile video. They reported a recognition rate of 87%. Segmentation can also be implemented with active contour using implicit and explicit shape representation. A recent method by [WCL11] proposed the use of constraint local models which jointly minimises the regularised misalignment error over predefined facial landmarks using a prior shape model and a local patch based landmark appearance model.

In the next section, a review of how to code a facial expression image

towards facial expression recognition is discussed.

2.3 Facial Parameter Coding Systems

Facial expression metrics need to be measured and represented in a consistent way in order to improve the discrimination of facial expression data. Various measurement systems have been proposed in the past. The most widely used are the Facial Action Coding System (FACS) [EF78], Facial Animation Parameter (FAPs) [DC10], the facial expression spatial charts and the maximally discriminative facial movement coding system (MAX) [IW79].

The most widely adopted scheme is FACS. This represents a fully standardized classification system of facial expressions. Due to its subjective nature, being coded by humans, FACS carried out on a face, could be biased and annotating images is time consuming. This makes the encoding of all the visually discriminative facial expressions on the human face possible. In the FACS framework, the occurrence of facial expressions is the combinations of elementary components called Action Units (AUs). Each AU corresponds to an individual face muscle group, which can be identified by code such as AU1, AU2, etc. There are about 44 unique AUs, each one of which reflects distinct momentary changes in facial appearance. The intensity of each AU is determined by action descriptors with the addition of a letter in the range A-E (A: trace,..., E: maximum, e.g. AU 12D) [KSMEK13].

Moving picture expert group (MPEG) is another defined facial animation coding system under the ISO/IEC 14496 (MPEG-4) standard introduced in 1999 [PP06]. The FAPs are a set of parameters that represent a complete set of facial actions combined with head motion, tongue, eye, and mouth control. In other words, each FAP is a facial action that deforms a face model from its neutral state. The FAP value shows the magnitude of the FAP thereby indicating the magnitude

of the deformation that is caused on the neutral model. For example, a small smile versus a big smile.

2.4 Facial Feature Extraction Methods

Accurate feature extraction is an important task in the process of FER [EF03]. Flaws in the process are likely to result in a biased recognition. Therefore, in order to evaluate facial expressions for emotion recognition, there is the need for the extraction of facial features that are specifically required for processing the data. Facial features can be divided into stable and transient. Permanent facial features such as, the lips, mouth, eyes or furrows that have become permanent due to age are stable facial features. Stable facial features can be deformed due to muscles movement to give facial expressions. Temporary facial movements or deformation appear mainly around the regions of the mouth, eyes and cheeks. These result when feelings or emotion are being expressed and can be categorised as transient facial features.

Existing methods of facial feature extraction for FER focus on extracting features from the whole face, an approach known as holistic [EF03]. Extracting specific facial regions of interest such as, the mouth, nose or eyes, is typically a local approach [EF03]. Whilst these two approaches can be used interchangeably, in reality, the holistic approach provides a complete picture of the facial information, sometimes, facial expression can significantly affect specific facial regions. Therefore, in such cases, the local approach which provides specific details and distinguishable information could be an option [DHS12].

Facial feature extraction methods can be divided into different categories based on the type of information they provide. In the wider research literature, the muscle/geometric based, model/appearance based, and motion based are the most frequently used [RAI12]. The combination of all three categories, the hybrid

methods is another category. Figure 2.2 shows a categorisation of facial feature extraction methods.

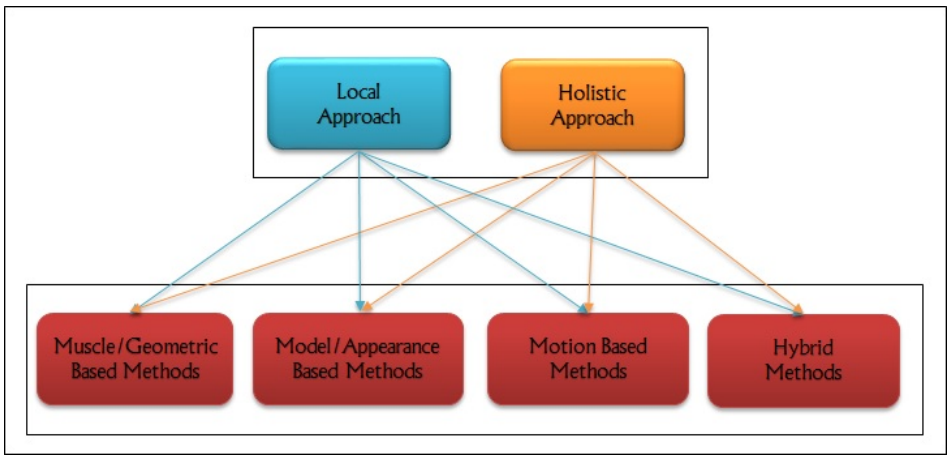

Figure 2.2: Categorisation of facial feature extraction methods

2.5 Facial Feature Representation

Facial feature representation is to derive a set of features from original face images to effectively represent faces [DBBB03]. A successful facial feature representation method should be able to obtain an optimal facial feature representation from the presented image data. This should minimise the within-class variations of the expressions, while maximising the between-class variations in order to perform robust and accurate classification [LPR14]. This could come with some challenges capable of reducing the classifier's accuracy and speed, such as the existence of a high number of extracted features, the quality and amount of data being analysed in the processing circles [LWY*13]. In essence, if inadequate features are used, even the best classifier could fail to achieve accurate recognition. Therefore, the classification performance highly depends on a proper feature selection/representation [DBBB03].

Sometimes techniques used for facial feature representation could result in a large bundle of features being produced [WLH*07]. This has the potential

of increasing the algorithms computational cost. Another kind of method to represent faces is to model the appearance changes of faces. These methods (dimensionality reduction) are used to project the high dimensional variables to a lower dimensional feature subspace. Holistic spatial analysis including Principal Component Analysis (PCA) [BS14], Independent Component Analysis (ICA) [LMMHM01], Linear Discriminant Analysis (LDA) [EC97], and Gabor wavelet analysis [SGM09], have been applied to either the whole face or to specific facial regions to extract facial appearance changes. Ideally a dimensionality reduction technique should be incorporated into every face analysis system. However, for FER, it can be avoided since the dimension of the extracted features depends significantly on the feature extraction method and this can only affect some of the classification methods [LLSC13].

2.5.1 PCA

PCA has been widely used for face and facial expression recognition. PCA performs dimensionality reduction by mapping data from a higher dimensional space to a lower dimensional space [BS14]. In performing FER based on the appearance approach, the pixel value constitutes the fundamental unit of information. The features may be extracted from a pixel set by either cropping or scaling and filtering. Thus, even at low resolution, the number of pixels in a face image is on the order of hundreds. Besides, many of the pixels present in a feature vector may contain little information which is useful for classification. One possibility is there could be cases where pixels located in certain regions of the face may not change in facial expressions, such that the corresponding feature vector coordinate becomes unusable. Another possibility is that a single pixel value in the feature vector might be completely dependent on other pixels. In other words, feature vectors could contain redundant information. In such cases, the classifier could be speeded up by removing the superfluous components [DHS12].

In order to get around the problem of high dimensionality, a linear combination of features is carried out to reduce the dimensionality [DHS12]. It is a method to linearly compress the features, by projecting the high dimensional data onto a lower dimensional space, while retaining as much as possible the variations present in the data set. PCA is one of the approaches used for finding effective linear information in order to transform the input variables into a new set of variables - the uncorrelated and ordered principal components, such that the first few components retains most of the variation information present in all of the origin. For example, an image vector with 65,536 pixels (256×256) might be projected into a subspace with only 100-300 dimensions. Several studies involving FER using the appearance based approach apply PCA prior to expression classification, for example [DBBB03] and [FL03]. Rudovic et al [RPP12], proposed an effective simultaneous FER of multiple emotions and their intensity estimation. PCA was used to construct a low dimensional input data representation in order to preserve discriminative information about various facial expressions of emotions and their intensities while being invariant to intra-and inter-subject variations.

Given that each feature vector $\mathbf{Z}$, is a column vector of M-dimensions or M-elements, consider a set of N_f feature vectors, which forms a feature matrix with one feature vector per column. Therefore,

$$Z = \{Z[m,n] \mid 1 \leq m \leq M, 1 \leq n \leq N_f \mid\} \tag{2.4}$$

PCA can be used to find a linear transformation through mapping of the original M *dimensional* feature space into L *dimensional* feature subspace, where normally $L < M$. The new vectors can be defined by:

$$y = \{y[m^{'},n] \mid 1 \leq m^{'} \leq L, 1 \leq n \leq N_f\} \tag{2.5}$$

and

$$y = w_{pca}^{T} z \tag{2.6}$$

where w_{pca} is the linear $M \times L$ transformation matrix, the superscript T denotes the transpose of matrix and the columns of w_{pca} are the L eigenvectors associated with the L largest eigenvalues of the covariance matrix c which is given by:

$$c = \frac{1}{N_f} \sum_{n=1}^{N^f} (z[m,n] - z_\mu[m])(z[m,n] - z_\mu[m])^T \tag{2.7}$$

and

$$z_\mu[m] = \frac{1}{N_f} \sum_{n=1}^{N^f} z[m,n] \tag{2.8}$$

where $z_\mu[m]$ for $1 \leq m \leq M$, is the mean of all feature vectors.

2.5.2 ICA

ICA is a generalisation of the PCA method, in which the goal is to find a linear representation of data so that the components are statistically independent [LMMHM01]. It is the most common method for generating spatially localised features in order to produce basis vectors that can make it statistically independent. Basically, the performance of ICA depends on the task to achieve, as well as the algorithm used to approximate ICA and the number of subspace dimensions retained. Two lines of different applications of ICA exist for data intended to be used for FER: ICA can be applied so as to treat images as random variables and pixels as observations, or the reverse, to treat pixels as random variables and images as observations.

To put the technique in perspective, Draper et al [DBBB03] used ICA to produce spatially independent basis vectors as architecture I and the use of ICA to produce statistically independent compressed images as architecture II. They showed that the FastICA algorithm configured according to ICA architecture II

yields the highest performance for identifying faces, while the InfoMax algorithm configured according to ICA architecture II was better for recognising facial actions [DBBB03].

2.5.3 LDA

Similar to PCA and ICA, LDA is a linear technique that projects the data onto a lower-dimensional vector space such that the ratio of the between-class distance to the within-class distance is maximised, thus achieving maximum discrimination [EC97]. The facial expression and analysis study, Wang, et al [WLL*10] used spontaneous visible images on front lighting using several typical methods, such as the eigenface (PCA), fisherface (PCA+LDA), the Active Appearance Model (AAM) and the combined AAM-based+LDA (AAM+LDA). They used PCA and PCA+LDA to recognise expression from spontaneous infrared thermal images. They claimed that LDA improves the rates of recognition in discrete dimensional expression recognition, while, it causes the performance to progressively worsen the emotion recognition category [WLL*10].

Recently, a FER study was conducted in [SAK*15]. A stepwise LDA (SWLDA) technique was employed for feature extraction and the hidden conditional random field (HCRFs) model was used for recognition. The SWLDA uses the partial F-test values for selecting the localised features from the expression frames. They gained a reduction in the within class variance and an increase in the low between variance among different expression classes [SAK*15]. Also, the authors claimed that the HCRF has the capability of approximating a complex distribution using a mixture of Gaussian density functions. They reported a weighted average recognition rate of 96.37% across four different data sets, which was significant compared with existing FER methods.

2.6 Facial Classification

Classification is the final stage of any facial expression analysis system [NPDlT08]. This often comes after any face detection and feature extraction. Thus, it is the process by which class labels are attributed to a set of facial measurements. Essentially, the classification process aims to correctly categorise the extracted features into different facial expressions of emotions. This can be achieved as facial expression interpretation or as facial expression recognition. When interpreting facial expressions, specific facial patterns are classified into mental activities or emotions. This involves directly associating the facial expressions with a predefined number of emotions or mental activity categories. Facial expression has been described at different levels [BT03]. The most widely used is the FACS, which is a human-observer based system developed to capture subtle changes in facial expressions. As mentioned in Section 2.3, FACS contains the most perceptible facial changes used to describe facial expressions by Action Units (AUs). Sample AUs extracted from the Cohn-Kanade database are shown in Figure 2.3.

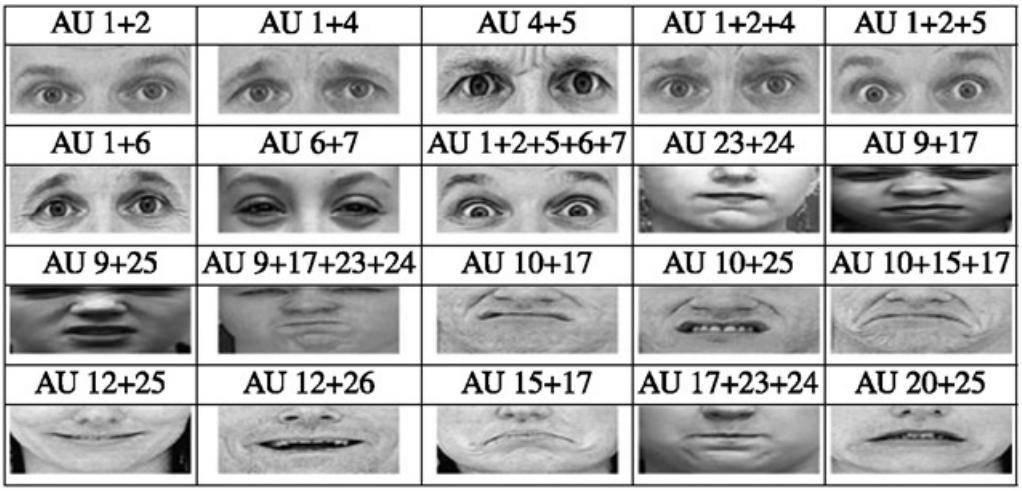

Figure 2.3: Sample action units from the Cohn-Kanade Database [TKC01]

The output of a facial expression classification module can either be the

recognised AUs or the six basic emotions. If the output are AUs, each of them gives an estimate of the probability of the input image consisting of the associated AUs as show in Figure 2.4

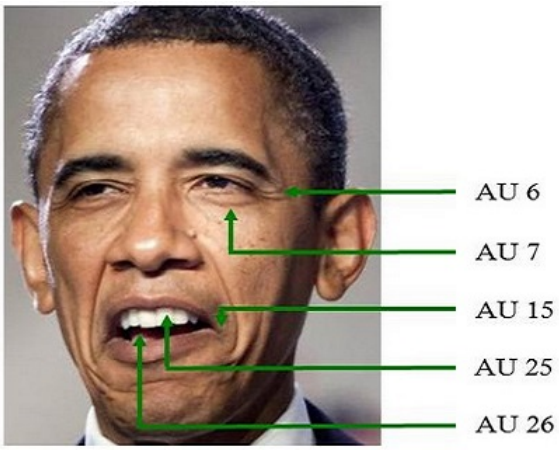

Figure 2.4: Facial annotation using Action Units from the FACS framework.

On the contrary, when the outputs are from the six basic emotions, only one output with the highest probability is selected. In general, the purpose of a classification process is correctly categorising the extracted features into the different facial expressions - a basic task for which the classifier has been applied. De la Torre and Cohn [DlTC11] divided classification techniques into supervised and unsupervised approaches. By convention, many FER systems designed to automatically interpret facial expressions requires labelled data for training. Such systems are based on a supervised approach, where classes are well defined in advance. There are systems defined based on facial expression dictionaries/rules which are used to translate facial actions into the different expression categories. Such systems operate an unsupervised approach. Training is done without labelled data, whilst classes are discovered. Most studies in this domain focuses on classification into the six basic emotions, whereas only few studies focused on detecting the non-basic affective states, for example, fatigue [FSYG10] or mental states such as frustration [RB09]. In the literature, the Hidden Markov Models

Table 2.1: Facial muscles description of FACS - AUs. [FE78]

Emotion	**Darwin's description**	**FACS, AUs Ekman's description**
Anger	Nostrils raised, mouth compresses, furrowed brow, eyes wide open, head erect	4, 5, 24, 38
Contempt	Lip protrusion, nose wrinkle, eyelid partial closure, eyes turn away, upper lip raised	9, 10, 22, 41, 61 or 62
Disgust	Lower lip turned down, upper lip raised, mouth open, out protruding lips, tongue protruded	10, 16, 22, 25 or 26
Fear	Eyes open, mouth open, lips retracted, eye brows raised	1, 2, 5, 20
Happiness	Eyes sparkle, under eyes skin wrinkle, mouth drawn back to corner	6, 12
Sadness	Mouth corner depressed, eyebrows inner corner raised	1, 15
Surprise	Eyebrows raised, mouth open, eyes open, lips protruded	1, 2, 5, 25 or 26

(HMMs), Dynamic Bayesian Networks (DBNs) and the Naive Bayes Classifiers (NBCs) are popular supervised learning methods [PP06]. Popular unsupervised learning methods are the geometric-invariant clustering algorithms, aligned cluster analysis (ACA) and the Active Appearance Models (AAMs) [PP06].

2.6.1 Emotions Interpretation

Two approaches used to interpret emotion interpretation from facial expressions data are message judgement and sign judgement. In message judgement, emotions or social signals are read out of the face in the same way as the observers' interpretation. With sign judgement, less emphasis is placed on the expression semantics, but more on higher level decision making. Sign judgement is widely used in psychology research, expression synthesis and affective computing [Coh07].

Generally, changes in facial muscle activities are usually very brief lasting

for only a few seconds, but are rarely more than 5s or less than 250ms [SSM12]. In order to track and interpret them, the location of the facial actions, the intensities and dynamics are very important. Facial expression intensities can be measured either by determining the geometric deformation of the facial features or the density of wrinkles present in certain facial regions. For example, anger is communicated by lowering the eyebrows and tightening of the lips, surprise is characterised by raising of the eyebrows and opening of the mouth, while smile is communicated by the magnitude of cheek and the raising of the lip corner as well as wrinkle displays [EF03].

In the research community, the interest has been to classify human emotions into some specific categories. This was led by the early research work of [EF71], in which theories that divides and categorise emotions are provided. Similarly, the early philosophy of mind [Plu01] put forward that all emotions can be categorised into basic classes (pleasure and pain). The philosophy of mind theory is described using the wheel of emotions [EF71]. This presents eight primary emotion dimensions arranged in pairs of four opposites - *joy vs sadness, trust vs disgust, anger vs fear and surprise vs anticipation*. Apart from the primary emotions, other emotions can be identified from their intensity differences and the mix of primary emotions. Figure 2.5 shows the wheel of emotion. Facial expression are one of the many correlates of emotions, since they are the most apparent ones. Humans are able to produce a lot of slightly varying sets of facial expressions, however, there is a small number of categories of distinctive facial configurations that almost every one associates with certain emotions, irrespective of their gender, age, cultural background and socialisation history. The categories of basic emotions are *anger, disgust, fear, happiness, sad and surprise* [CB03]. Based on research findings almost everyone can produce and recognise the basic facial expressions of these emotions has led to the assumption that they are universal emotions [TKC05].

Emotions are brief conscious experience brought about by the mental

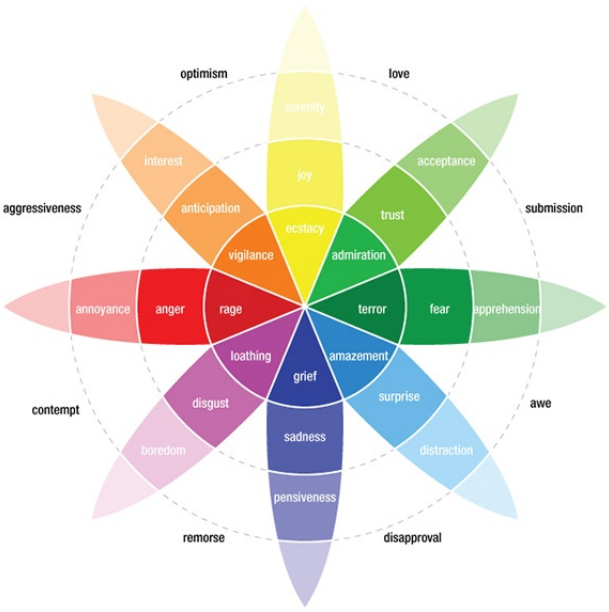

Figure 2.5: Plutchik's wheel of emotions [Gol14]

activity and a degree of pleasure/displeasure. They are closely related to the physiological and psychological levels of arousal to specific emotions. In neurobiology, emotions are taken as complex action programs triggered by the presence of certain external or internal stimuli, of which facial expressions is one of the elements [Mür15]. Computer vision systems can be used to extract and interpret emotions from facial expressions by studying specific facial regions or geometry of the facial structures. Emotion classification can lead into the identification of basic emotions and the estimation of psychological status of an individual. The SVMs classifier will be described in the next section.

2.6.2 SVM

SVMs are a primary classifier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels [CST00]. Technically, SVMs supports both classification and regression problems and can also handle multiple continuous categorical variables.

This thesis is more concerned with the classification problem. Given labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. SVMs are thus based on the concept of decision planes that define decision boundaries. A decision plane is one that separates a set of objects having different class memberships.

Improving a classifier's effectiveness has been an area of in-depth machine learning research [CVL07]. A lot of research works in this domain have been successfully applied to pattern classification problems, particularly facial expressions classification, for example [NPDlT08, CVBM02]. The SVM decision hyperplane can be defined by an intercept term, z and a decision hyperplane normal vector $\mathbf{d}$, which is perpendicular to the hyperplane [DHS12]. The vector is referred to as the weight vector. Let χ be a set of N_s labelled training set

$$\chi = (X_i^f, c_i) \mid c_i \epsilon \{-1, 1\}, 1 \leq i \leq N_s \tag{2.9}$$

where X_i^f are the training samples and c_i is the class label X_i^f. The linear classifier $(h(X^f))$ typically try to find a decision function given by:

$$h(X^f) = sgn(\langle \mathbf{d}, X^f \rangle + Z) \tag{2.10}$$

and

$$sgn(\tau) = \begin{cases} 1, \tau \geq 0 \\ -1, \tau < 0 \end{cases} \tag{2.11}$$

where $\mathbf{d}$ is the decision hyperplane normal vector and z is intercept term. $h(X^f)\epsilon\{-1, 1\}$ yields a label (-1 indicates one class and +1 indicates the other class) for an unseen example X^f. The SVM linear classifier relies on a dot product between data point vectors.

Let $K(X_i^f, X_j^f) = X_i^{f^T}, X_j^f$, the SVM classifier h_{svm} is given by:

$$h_{svm}(X^f) = sgn(\sum_{i=1}^{N_s} \alpha_i c_i K(X_i^f, X^f) + z) \tag{2.12}$$

where the α_i are Lagrange multipliers of a dual optimisation problem.

If in the optimal solution, all of the α_i are shown to be non-zero, then that is those training points nearest in neighbour to the hyperplane. This is referred to as the support vectors. This brought about spareness in the solution and gave rise to efficient approaches to optimisation [WMC*00]. Once a decision function is reached, the classification of the unseen examples (X^f) are checked to ascertain where on the hyperplane these lies. Technically, the SVMs carry out implicit embedding of data into a high dimensional feature space, where the separation of data that are only separable with nonlinear rules will be carried out in input space [MEK03]. In order to achieve this, the learning algorithm is formulated to make use of the kernel functions, which will allow for the efficient computation of the inner products directly in the future space, without the need for explicit embedding. A kernel function (K) in the expanded future form, corresponds to a dot product. The Equation 2.13 describes the kernel function

$$K(X_i^f, X_j^f) = \langle \phi(X_i^f)^T . \phi(X_j^f) \rangle \tag{2.13}$$

where ϕ is a nonlinear, mapping function which embeds input vectors into feature space. Using a kernel function, SVM can be used as an alternative training method for polynomial, radial basis function (RBF) and multi-layer perceptron (MLP) neural network classifiers in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-convex and unconstrained minimisation problem as in standard neural network training. The native RBF is usually given by a Gaussian distribution

$$K(X_i^f, X_j^f) = exp[\frac{\|X_i f - X_j^f\|^2}{2\sigma^2} \tag{2.14}$$

where $\| \bullet \|$ is the Euclidean distance between two vectors. In 2.14, the output of the kernel is dependent on the Euclidean distance of X_j^f from X_i^f. By convention, the output will be one support vector and the other the testing data point. The centre of the RBF becomes the support vector and ϕ is used as the variable to determine the area of influence of the support vector over the data space [PFW*11]. A ϕ with large value will be an advantage to give a smoother decision surface and more regular decision boundary. With a larger ϕ, the RBF will allow a support vector to have a strong influence over a larger area. Also, a larger α value has the potential of increasing the value of α_i in the Equation 2.12. Generally, when a support vector influences a larger area, all support vectors in the neighbouring area will also increase in α_i value in order to counter the influence, so that all the α_i will reach a balance at a larger magnitude. On another perspective, a larger value of α_i will reduce the number of support vectors, with a larger space covered - fewer α_i will therefore be needed to define a boundary.

2.7 Facial Expression Recognition

Ekman et al [EF71] carried out an important step in facial expression analysis. They identified six universal facial expressions that corresponds to specific emotions which are expressed in the same way independent of age, gender or cultural affiliations the world over.

Facial expression analysis can be carried using two popular classes:

1. Analysis that determines a single facial action (i.e. raising a single eyebrow, closing the eyes, opening the mouth, etc). This approach usually proposes a single facial action or a set of states for facial components of which some are semantically exclusive and some are independent. In some instance, the state of the eyes can be presented as wide or closed, sometimes, the two states can be combined with another state, such as opened mouth. The description

of the facial expression for this approach uses the FACS framework, which describes facial expressions by activation of the AUs in the face. Thus, FACS [FE78] is a technique for the measurement of facial movement, where each AU is a single intra-face movement and the contraction of specific facial muscles. FACS and AUs have been described in Section 2.3. Sample facial annotation using FACS and AUs framework is presented in Figure 2.4.

The advantage of this technique is its flexibility. Combining activated action units are indicative for many states of mind apart from emotions, for example, confusion, pain, fatigue etc. On the contrary, manually labelling images/image sequences is a tedious task and, much effort is needed to analyse this task [ZPRH09]. A non person-specific approach has been proposed in [BLF*06], where an image-based system is developed to recognise the activation of 27 different facial AUs by applying Gabor-wavelets to the facial region and utilising SVMs and AdaBoost for classification.

2. The other approach is facial expression analysis that determines a complete facial expression. This is usually done with a specific semantic (angry, happiness, pains, etc). Here, predefined facial expressions rather than a single facial action is dedicated to classifying the six universal facial expression of Ekman et al [EF71]. In this approach, facial expression can be considered static, that is, as a fixed face state, or dynamic, that is, as intra-face movement. In the work of [WLF*09], an image-based approach was followed to determine smiles. The system relied on AdaBoost and SVMs to determine the facial expression from convolution of image data with Gabor energy filter.

The early attempts of research in facial expression analysis built systems that are restricted to certain conditions, such as, face images being captured at frontal or profile view, capturing under controlled lighting conditions, etc. Systems

were thus built to know the location of the face of facial landmarks. Also facial expression analysis were restricted to acted and exaggerated basic emotions. To date, though, remarkable progress has been made in the field of facial expression analysis [SI92]. In addition, the physiognomies of faces vary considerably between different people. This is attributed to the differences in age, ethnicity, gender, facial hair, cosmetic products and occlusions such as glasses, scarf and hair. As a result, the automation of facial expressions task becomes more complex. Also faces may appear different because of pose and changing light. Therefore, in order to achieve a meaningful result, these variations have to be addressed at different stages of an automatic facial expression analysis pipeline.

2.7.1 Facial Expression Architecture

Generally, a duality of task exist between face recognition (FR) and facial expression recognition (FER). In the literature, similar structure and processing techniques are often used for both tasks. In both tasks, there is the consideration of what is wanted for FR or FER. Thus, faces convey other information such as the identity of a person in addition to expressions. Personal identity information conveyed by the face is an unwanted source of variability for FER tasks. In line with that, the variability arising from facial expression is unwanted in face recognition. Therefore, the uniqueness of the face is the central criteria for the recognition task. In order to adhere to the classical pattern recognition model, FER tasks usually take the form of a sequential configuration in the following way: image acquisition, pre-processing, feature extraction, classification and sometimes post-processing. The individual stages of the FER architecture is presented in Figure 2.6

Based on the spatial nature of the human face we mostly perform FER tasks with two vital aspects: feature extraction and classifier design. In previous work feature extraction has been used to estimate the displacements of feature points or optical flow analysis has been used to model muscles activities [TKC05]. However,

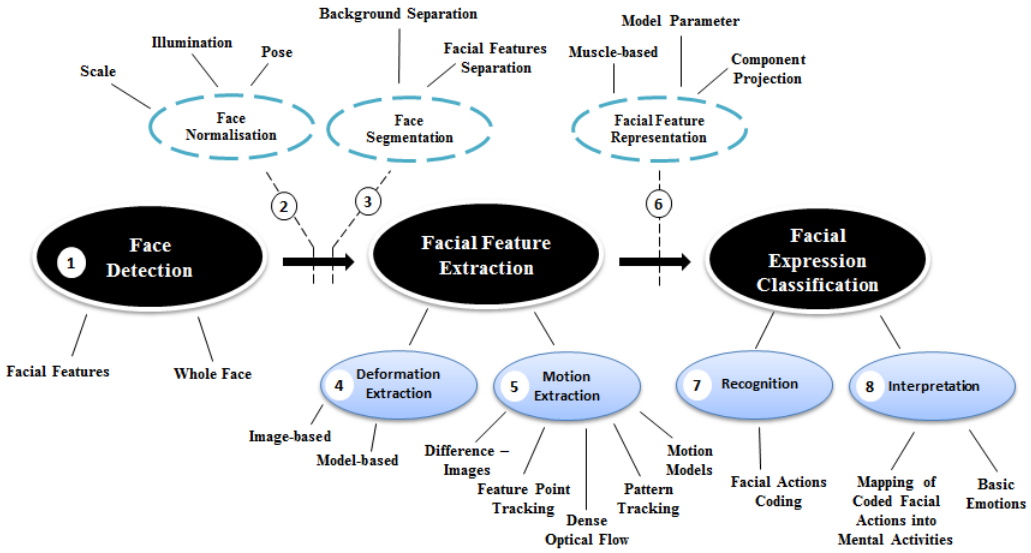

Figure 2.6: Sample FER Architecture

the estimation of optical flow is prone to distortion by the nonrigid motion and changing lighting conditions. This can lead to inaccuracy of image recognition. Also facial geometry has been used in facial representations, where shapes and locations of facial components are extracted [PP06]. In [ZLSA98], the authors used the geometric position of 34 fiducial points as facial features to represent facial images.

For the purpose of processing, FER can be classified as local or holistic. Local recognition typically involves an ensemble of feature extractors/classifiers, together with a combination unit. Unlike local recognition, with holistic, the whole face provides a single input for the recognition system. Most of the sequential configurations when using the classical pattern recognition model for FER listed above have been reviewed earlier in this chapter. Image acquisition and pre-processing are described here. Two categories of images have been adopted for FER: static image and image sequence (video). A still image is typically an 8-bit intensity image or three-channel RGB with 8 bit per channel. Similarly, an image sequence is a collection of images related by time or spatial location, the set

of these images can be represented with a brightness function $g(x, y, t)$ referred to as temporal image sequence.

2.8 Existing Databases

Several databases have been proposed which contain facial images for computer vision purposes, including [LAKG98, DGLG11a]. These databases consist either of single images or image sequences, mainly captured in laboratory conditions. The information on the databases content is usually provided by the database authors. Depending on the purpose of the database, this is called "meta-data" and includes the person's identity, facial expression, action unit activation, annotated facial components or landmarks.

To be able to validate the facial expression based on image-data under harsh lighting conditions that will be discussed in this thesis, we need a suitable set of data. There have been a number of freely available databases in the public domain, none of which addresses successfully all the FER problems intended to be solved in this thesis. Table 2.2 lists a number of important characteristics of the databases. In the table we have included information on 10 facial expression databases that have been used in the literature on FER. A number of important characteristics are also listed: the number of subjects, videos and images contained in the database as well as whether the database is publicly available. In the following Sections, the most popular and those used in this thesis will be discussed. Two publicly available databases: the Japanese female facial expression (JAFFE) and the Static facial expression in the wild (SFEW) were chosen. The other database used is the HDR dataset, which was created in this thesis for the purpose of providing the facial expression recognition community with a dataset for developing and validating algorithms for facial expression analysis. This will be discussed in Chapter 3.

Table 2.2: Characteristics of facial expression databases

Name	Expression description	No. Subjects	Characteristics	Views
AR [Martinez & Benavente, 1998]	3 basic emotions	126	Glasses, scarf, different illumination	Frontal
AT&T (formerly ORL [Samaria & Harter, 1994]	2 basic emotions	30	Glasses, varying lighting	Frontal
BU-3DFE [Yin et al., 2006]	6 basic emotions	100	4 levels of intensity (3D)	Frontal, 45, -45
Cohn-Kanade [Lucey, Patrick et al., 2000]	7 basic emotions	97	Uniform lighting	Frontal, 30 degrees
JAFFE [J. et al., 1998]	6 basic emotions	10	Uniform lighting	Frontal
MMI [Pantic et al., 2005]	6 basic emotions	90	Glass, facial hair	Frontal, profile
SFEW	6 basic emotions	95	Non-uniform lighting	Frontal
PIE [Sim et al., 2002]	4 expressions	68	43 lighting conditions	Frontal
NimStim [Tottenham et al 2009]	6 basic emotion	31	Uniform lighting	Frontal
MultiPIE [Gross et al., 2007]	6 basic emotions	43		Frontal

mJAFFE: The JAFFE [LAKG98] facial expression database has been extensively used for facial expression analysis. Each female has two to four samples for each expression, totalling 213 greyscale facial expressions images from 10 Japanese female actresses. It consists of six basic emotions (*anger, disgust, fear, happiness, sadness, surprise and neutral*). The expressions expressed by each picture were subjectively tested on 60 Japanese volunteers. Each image is of size 256×256. Figure 2.7 shows a sample image from the JAFFE database.

Figure 2.7: Sample Six Basic Emotions from the JAFFE database. *Starting from top, we have angry, disgust, fear, happy, sad and surprise*

Since the JAFFE database is a standard database that has been accepted and used on a number of occasions for validating facial expression experiments. We decided to create a modified version of JAFFE called *mJAFFE* (modified JAFFE) with a Photoshop script by making: shadow, high contrast, low contrast, overexpose and underexpose lighting conditions. This is in order to make an artificial lighting conditions close to the HDR database, for the purpose of validation the FER experiments. Figure 2.8 shows a sample image from the mJAFFE database.

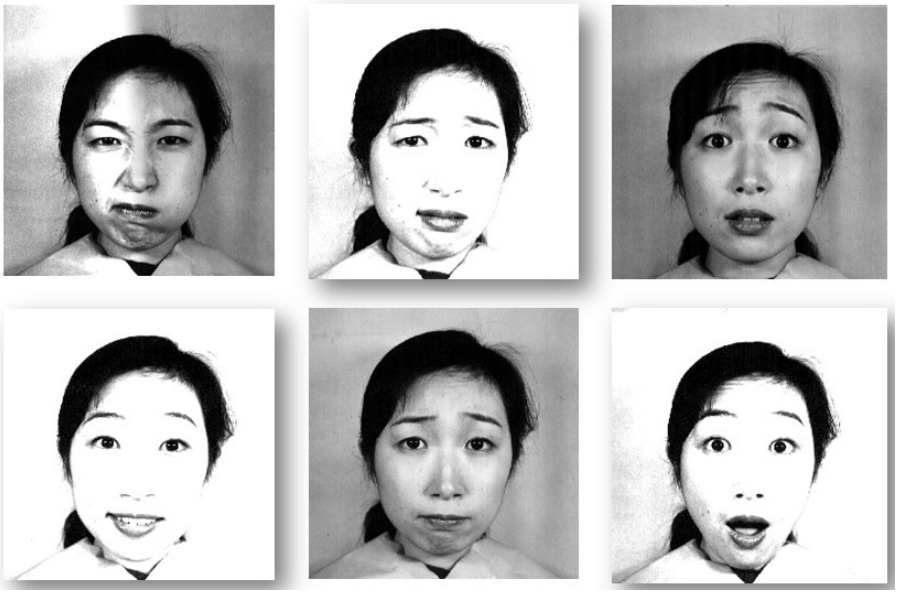

Figure 2.8: Sample Six Basic Emotions from the mJAFFE database. *Starting from top, we have angry, disgust, fear, happy, sad and surprise*

SFEW: The SFEW [DGLG11a] dataset was extracted from frames from Acted Facial Expressions in the Wild (AFEW) database. The database was collected with close to real world lighting conditions, with different head poses, large age ranges, different face resolutions, occlusions and different focus. There are 95 subjects and a total of 663 well-labelled usable images. Figure 2.9 shows a sample image from the SFEW database.

Figure 2.9: Sample Six Basic Emotions from the SFEW database. *Starting from top, we have angry, disgust, fear, happy, sad and surprise*

Image Acquisition

In image processing, image acquisition usually involves retrieving images from a source that is automatically capturing images [CB03]. For example, real-time image acquisition creates a stream of files that can be automatically processed, queued for later work, or stitched into a single media format. There are methods of image acquisitions in image processing that actually uses acquisition devices, for example, using digital cameras to acquire images to build accurate models of different scenes.

Ultimately, in an image processing task, image acquisition is always the first step because, without an image, no processing is possible. The image that is acquired is completely unprocessed and is the result of whatever hardware was used to generate it, which can be very important in facial expression process to have a consistent baseline from which to work. One of the purposes of an image acquisition process is to have a source of input that operates within such controlled and measured guidelines that the same image can, if necessary, be nearly perfectly

reproduced under the same conditions so that inconsistency factors are easier to locate and eliminate.

In this thesis, due to the type of images used (those under harsh lighting conditions), it is important to relate an image to the scene light from which it was captured. In practice, often, facial images are subjected to changes in viewpoints, lighting conditions and expressions, wherefore providing clear, highly detailed images still needs more improvement. For instance, one of the key elements in getting a good image is how well the camera can cope with the scene changes in lighting and weather conditions. Even with the best camera on the market, image variability is still an issue. The first issue is how to choose the best features to represent the face in order to deal with the facial variability.

2.9 Facial Expression Recognition Methods

There have been several methodologies used to interpret facial expressions within computer science. These can be grouped under audio based methods and vision based methods. In addition, there has been some research which have used a multimodal signal (combination of audiovisual signals) to achieve better results, for example [RBFD03, ZHR*07, CMK*06]. In this thesis, the focus is on the vision only based methods, that is, expressions a determined from a face without a sound being produced. Vision based methods for analysing facial expressions operate on images and image sequences. Many studies have adopted the popular feature engineering technique to perform FER tasks using data affected with harsh lighting conditions, ranging from the pre-processing of the training and testing images to normalisation of the image lighting, the removal of lighting or the equalisation of lighting effects. As discussed previously (Section 2.2), there are many algorithms aimed at improving lighting conditions in specific regions of the images. There are several types of problems for which these algorithms are proposed, such as if

an image is unevenly illuminated and another portion is in overly bright sunlight. The idea is to improve the contrast in both parts of the image without creating an unnatural result.

The traditional approaches [LHK05, TT10] for performing FER tasks using data affected with harsh lighting conditions can be categorised in the following ways:

1. Direct appearance-based methods. In this approach, training samples are usually collected under different lighting conditions and used directly to learn a global model of the collected illumination variations samples. This is done without performing lighting pre-processing on the images. Examples of this method are the linear subspace and the manifold model [TT10]. This methods classifies based on the lighting variations present in the new images. It has been observed that direct learning using this method makes few assumptions on the results, and it requires a large amount of training samples and an expressive feature set.

2. The normalised-based methods performs a canonical form of reduction on the training images in order to suppress the lighting variations. A typical example of this method is the histogram equalisation, where the facial lighting are equalised. There are other customised methods which exploit the idea that a natural distribution of incoming illumination is predominantly of low spatial frequencies and soft edges, for example [LHK05]. The high frequency information in the image is predominantly signal, which is an intrinsic facial appearance. For examples, the Multiscale Retinex (MSR) method was used in [RJW04] to cancel much of the low frequency information, where the image is divided by a smoothed version of itself. These methods which has also been used in other studies for the purpose of dealing with image lighting variations, have been shown to be quite efficient

but limited in their ability to handle non-uniform lighting variations.

3. With feature-based methods, a direct extraction of the image's insensitive lighting feature sets is carried out, including the geometric features, image derivative features, Local Binary Patterns (LBP), Gabor wavelets and local autocorrelation filters [TT10]. Whilst a feature based approach offers promising improvement on raw grey scale images, their ability to deal with real-world complex lighting variations is still limited. It has been reported [CBJ00] that a complete illumination invariants is difficult to attain, however, finding representations that can overcome the most common classes of natural lighting variations can overcome the limitations

2.10 Summary

This chapter presented a review on faces, facial expression and emotions. It also covered general methods and a pipeline for performing facial expression analysis particularly focused on harsh lighting conditions. In addition, the databases used for validation purposes in this thesis, were introduced. The background presented in this chapter forms the foundations for understanding the remaining chapters.

Chapter 3

High Dynamic Range Imaging Review

"Sometimes it's necessary to go a long distance out of the way in order to come back a short distance correctly." Edward Albee

3.1 Introduction

The image lighting conditions discussed earlier in Chapter 1 have been identified as a potential problem for FER systems. Harsh lighting can result in the permanent loss of information in image pixels [WLH*07], for example, the changes introduced by lighting variations are often larger than the differences between individual facial expressions [FN09a]. Drbohlav and Chantler [DC05] reported that intra-class variations due to harsh lighting conditions is greater than the inter-class variations on different individual. In particular, when the perception of an image changes due to the effect of harsh lighting conditions, this has the potential to affect FER performance.

Based on these, one of the aims of this thesis is to explore the possible benefits to be gained by utilising HDR methods for solving FER problems.

The dynamic range in a scene is the difference between the brightest and darkest areas of that scene [BADC11]. A High Dynamic Range (HDR) image can be generated either synthetically or acquired from the real world. HDR imaging was developed in order to capture, record and process the full range of light in a scene so that it could be reproduced as a photograph or image [KSZ*14]. Neither computer screens nor paper can display the full dynamic range that is present in the real world. As a result cameras were not designed to capture even close to such a range. HDR has the capability to represent a large luminance variation in either image or video signal, that is from very dark values $(0.0005cd/m^2)$ to very bright values $(> 1000cd/m^2)$. Existing systems, referred to as Low Dynamic Range (LDR) or Standard Dynamic Range (SDR), support luminance values only in the range of $0.0002\ to\ 100cd/m^2$. HDR creates brighter whites, darker blacks and colours to better match images we see in the real world. HDR allows a full range of lighting in a real world scene to be captured, stored, transmitted and displayed comparable to what the Human Visual System (HVS) can see at any level of adaption [BADC11]. The HVS is capable of adapting to various changes in lights that vary approximately by 13 orders of magnitude ranging from a starlit night $(10^{-5}cd/m^2)$. In comparison, an LDR system, can capture and display a dynamic range of 3 orders of magnitude at most.

LDR images are thus not suitable for capturing images particularly in scenes where typically harsh lighting exist and there is the need to capture all the light. LDR images are typically clipped in the range $0 - 255$. HDR images, on the other hand, use floating point numbers. The floating point data precision allows HDR imaging to represent real-world luminance values. This is sometimes referred to as scene referred data.

3.2 Dynamic Range Terminology

The dynamic range in photography is the number of stops between the darkest part of an image where you can still resolve detail and the lightest part. Digital Single-Lens Reflex (DSLR) cameras generally have about 11 stops of dynamic range at low ISO values, and point-and-shoot cameras a stop or so less [BADC11].

Dynamic range can be computed for different purposes. For example, in imaging, this is the difference between the brightest and darkest regions of details present/captured within a single image or on a display. The HVS can perceive about 14.5 stops at once, but can perceive greater ranges through adaptation. In a camera, it is the ratio of light saturation to noise or more precisely, the ratio of the intensity that saturates the camera to the intensity that lifts the camera response is one standard deviation above camera noise flow [mel]. In a display, dynamic range is the ratio between the maximum and minimum intensities emitted from the screen. In this thesis, our focus is on the dynamic range of a scene.

Dynamic range may also be termed contrast ratio and can be represented using the ratio notation, e.g 100 : 1. Contrast ratio (CR) is expressed as:

$$CR = L_{max} : L_{min} \tag{3.1}$$

where L_{max} and L_{min} are the maximum and minimum luminance values. Black is not considered as the darkest colour, as the function would contain division by zero, therefore the next smallest value is used. Figure 3.1 illustrates the concepts of CR, showing eleven different levels of brightness with corresponding values and the difference between the consecutive blocks represents the smallest change in brightness. The back part is not included in the calculation.

The luminance of starlight is around $0.001 cd/m^2$. A sunlit scene is around $100,000 cd/m^2$. The luminance of the sun is approximately $1,000,000,000 cd/m^2$

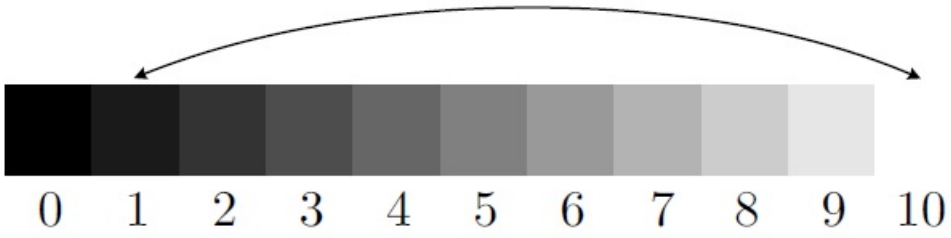

Figure 3.1: Example of 10 : 1 CR.

[BADC11]. Dynamic range can also be measured in the logarithmic unit stops, which represent doubling of light intensity, and can be used to plot the luminance to easily represent the different values. The log base 10 of the luminance is represented in a scanline in Figure 3.2. Going from *0.1 to 1* is the same distance as going from *100 to 1000*.

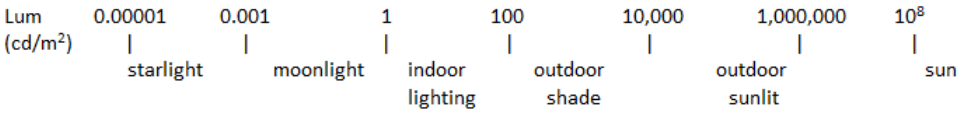

Figure 3.2: Scanline representing luminance. For example, for a scene showing the interior of an office with a sunlit view outside the window will have a dynamic range of approximately *100,000:1*

.

In scenes with changing light, the dynamic range may usually be a problem. For example, capturing an image with a standard camera on a sunny day from inside a building looking outdoors, could pose difficulties to accurately capture both the scene indoors and out doors simultaneously. The picture can end up with under-exposed or over-exposed pixels.

In a *well exposed* scene on the other hand, for instance, in Figure 3.3, in the facial region, there are no exposure concerns in bright and dark points, and thus dynamic range is not a problem. Comparing the image in Figure 3.3 with Figure 1.8, it can be seen that the images in Figure 1.8 contain both the bright and dark regions, which means that, the camera can either capture the bright or dark areas,

but not both. To capture both, one option is to bracket the exposures as described in Section 3.3. A truely high dynamic range image by definition, should normally be

Figure 3.3: Image captured in a *well exposed* scene. No difficulty with dynamic range.

able to capture all the intensities of light visible to the human eyes or more. There is no defined boundary/threshold which can be used to split between high and low dynamic range as, this depends on the subjective nature of human interpretation or judgement. There are no available DSLR camera sensors that can directly support HDR capture of at least > 16 stops at the moment. However, different solutions have been proposed and adopted to overcome this, such as capturing multiple exposures and merging them into a single image.

The HVS senses light nonlinearly, for example, when we increase the contrast of an image by 50 from $150 : 1$ to $200 : 1$ this impacts perception significantly, but when the same increase is made from $10,000 : 1$ to $10,050 : 1$, this may not be noticed. Based on this, fractional representation as shown in Figure 3.1 is not intuitive when discussing dynamic range. However, the exposure value measure is preferred by most researchers in image processing [AT12]. Exposure Value (EV) is the term used to describe the amount of light that is passing through the lens and hitting the sensor. The EV measure corresponds to doubling and

halving the amount of light - it follow a base 2 logarithmic scale. That means, HVS response to light is approximately logarithmic, which makes the EV scale a more natural measure of dynamic range. EV could be termed *stop*. *f-stop* is used to denote the photographic aperture size, it is sometimes synonymously used as EV to represent dynamic range as its rate of change is the same [AT12].

3.3 High Dynamic Range Capture

One technique for creating an HDR image is to use a set of images from the same imaging device (e.g. camera) with different exposures. That is, a number of LDR images with different exposure time are used. With a short exposure time, the camera sensor will have less time to saturate so the image will appear dark. Therefore, low exposure is used with bright areas of the scene. The longest exposure time is chosen so that the sensor will fully saturate and as a result, the image appears bright [VDJ10]. An exposure X_{ij} is defined as the irradiance E_i at the sensor multiplied by the exposure time Δt_j.

$$X_{ij} = E_i \Delta t_j \tag{3.2}$$

where i denotes the particular pixel or sensor location and j the exposure.

Thus in order to get around the limitations of LDR images, a still-image camera is used to take a number of images in quick succession at different exposure times. These are then being merged together to create a HDR image. Thus, HDR imaging is relatively easy to achieve in a still-image camera, as only a few single static images are needed to create the HDR image.

Figure 3.4 shows a set of 7 bracketed images with each exposure *11-stops* apart in luminance. This means each exposure is properly exposed for a different region in the scene. The final merged HDR image shown in Figure 3.5 contains details in both dark and light regions. Note this image has been tone mapped for

Figure 3.4: Seven bracketed LDR images captured under harsh lighting conditions with shadow on one side of the face.

printing in this thesis.

Figure 3.5: HDR image created from merging the LDR images above.

There are several algorithms used for merging LDR images, for example, Debevec and Malik$'$s method [DM08]. The merging process can also be automated as done by Spheron HDR VR [KA]. This can capture still spherical images with a dynamic range of $6 \times 10^7 : 1$. HDR images/videos, however occupy an estimated four times the amount of memory of uncompressed LDR images. Hence, this has a huge effect on both storage and transmitting, therefore, efficient representations of floating point numbers is an important part of HDR imaging. Recently, many conventional compression algorithms such as JPEG and MPEG have been extended to handle HDR images and videos [DM08].

After HDR images have been created, they need to be visualised. However, with current LCD monitors, the images/videos do not fit the dynamic range,

which is around 200 : 1. To display HDR content, it has to be further processed by dynamic range compression, known as tone mapping [BADC11]. Monitors capable of visualising HDR content were proposed by Seetzeen et al [SHS*04].

3.4 Tone Mapping

Tone mapping operators (TMOs) reduces an HDR image so that it can be rendered onto a LDR monitor in a manner that maintains the relative luminosities and reproduces the overall effect of the original scene, in a way that clamping effects occurring through the use of standard graphics output devices is avoided [DD02]. Within the past decade, researchers have developed many algorithms that will take a set of luminance values from a larger, unrestricted range to a much smaller, restricted range.

A good TMO could be described as one which produces an image that is perceptually similar to the original scene. Most of the algorithms operate on the luminance of an HDR image, since luminance largely governs our contrast perception [Aky12]. Tone mapping methods can either be global (also called spatially invariant) or combined with a local processing (also called spatially variant), modelling either only the global adaptation, or the global and local adaptation of the HVS.

3.4.1 Global Tone Mapping Operators

A global TMO uses a monotonic mapping curve to independently transfer real-world lighting and colour to the display. The algorithms apply the same function to all pixels of the image, that is one input value results in one and only one output value. For example, such a TMO can be a power function, a logarithm or a sigmoid curve across all values in the image of a function f that is image-dependent [DBDQ10]. The goal is to approximate the HVS's non-linearity,

to compensate for the display characteristics, or to render visually more appealing images. Global tone mappers tend to be more restricted in the aesthetic quality of the image they produce due to low levels of relative contrast between objects. Eilertsen et al [EWMU13] showed that a global strategy is, however, unable to capture important local transitions that may need to be preserved in order to maintain an overall level of local contrast corresponding to the original HDR input. In particular, logarithmic function have been applied in different ways [DMAC03] to luminance to increase contrast and brightness for the display of HDR images, on the low luminance values while compressing the higher luminance values.

Logarithmic TMO

A logarithm function is often used to approximate the non-linear encoding of the HVS. Thus, in the log-encoded image, equal steps in log-luminances correspond to equal visual sensations. This enables a perceptually uniform quantization where the perceived difference between two digital code values remains constant over the digital code value range. Such a logarithm function is used in the Retinex model of colour vision [EWMU13].

Gamma

Display devices have a non-linear relationship between input voltage and display luminance. This non-linearity is described by a power law and is commonly called gamma, which is referred to as the numerical value of the exponent. The output of a monitor can be modelled as follows:

$$L = V^{\gamma} \tag{3.3}$$

where V is the input voltage, γ is the gamma value of the display and L is the luminance produced at the screen.

In order to display corresponding luminance to those of the capture scene, the non-linearity has to be inverted. To achieve this, each colour channels of an input image I is processed as follows:

$$I_c^{'} = I_c^{\frac{1}{\gamma}} \tag{3.4}$$

where c denotes one of the RGB colour channels of the input image I, and $I^{'}$ is the gamma corrected image. The γ value depends on the monitor. A common average value of gamma is 2.2.

On top of compensating for the display non-linearity, an advantage of gamma encoding is that it approaches the functions described above in a way that models the HVS non-linearity. Thus, a gamma-encoded image is also approximately perceptually uniform [KYJF04].

Gamma Correction with Adaptation on the Image Key

In addition to compensating for the output medium non-linearity, we may also want to improve the reproduction of an image depending on its content. An image can be characterised by its dominant tones, which is called the key. The key of an image indicates whether it is subjectively light, normal or dark, thus [KYJF04] approximated it by the log average luminance $\overline{\wedge}$:

$$\overline{\wedge} = \exp(\frac{1}{N}\sum_{p\in I} log(\in + \wedge(p))) \tag{3.5}$$

where p is a pixel in the image I whose luminance channel is given by $\overline{\wedge}$, N is the number of pixels, and $\in$ is a small value to avoid singularities caused by the presence of black pixels.

The gamma exponent value may be adapted to the key of the image to render more pleasing images. However, when rendering a low key image, it is

desirable to carry out gamma correction with a greater gamma value to improve detail visibility in dim areas.

3.4.2 Local Tone Mapping Operators

Local tone mapping algorithms apply different functions for different spatial pixel positions. Instead of applying the same tone curve to the entire image, they adapt each pixel individually based on the surrounding pixels. That means, one input value can result in more than one output value depending on the pixel position and on surrounding pixel values. Local TMOs consider pixel neighbourhood information in the mapping processing for each individual pixel, which simulates the adaptive and local property of colour vision called Retinex. The goal of Retinex is to recover the perceived colours from the captured scene radiances [MMK08]. They improve the quality of the tone mapped image over global operators by attempting to reproduce both the local and the global contrast. Local tone mappers regard the area around a given pixel in order to best utilise the available dynamic range, otherwise halos around edges can appear. However, Halos are sometimes desired when attention needs to be given to a particular area [BADC].

To preserve local contrast, Chiu et al [CHS*93] proposed to preserve local contrast, where the TMO is used to scale the world luminance with neighbouring pixels average. This is defined as:

$$L_d(x) = L_w(x)s(x) \tag{3.6}$$

s(x) is the function for scaling applied to compute the local average of the neighbouring pixels, which is defined as:

$$s(x) = (k(L_w \otimes G_\sigma)(x))^{-1} \tag{3.7}$$

G_σ is a Gaussian filter and k is a constant that scales the final output.

3.5 HDR Facial Expressions Database

Sometimes faces can appear very different due to a number of reasons. This can be attributed to three significant reasons: the pose (the angle at which the face is viewed), the lighting conditions at the time, and the facial expression (whether or not they are being sad or disgusted). Amongst these, facial lighting conditions is of interest to this study and this led to the need for a database consisting of subjects with a variety of facial expressions under different harsh lighting conditions. Although, other facial expression databases exist with different lighting conditions [SBB02], these are LDR and not under harsher lighting conditions.

3.6 Summary

This chapter has introduced the main concepts of dynamic range: low dynamic range and high dynamic range, capture, storage, and the issues with displaying HDR images. Tone mapping operators were also discussed. The ability of HDR images to capture the full range of detail in a scene is a key feature which will be explored throughout this thesis.

Chapter 4

Research Methodology

"Life is a Mathematics equation. In order to gain the most,
you have to know how to convert the negatives into positives."

4.1 Introduction

This chapter details the discussions on the research methods used in the thesis, it discusses the adopted techniques and the steps taken in order to develop an understanding of whether high dynamic range (HDR) imaging can be useful for improving facial expression recognition (FER) system. Since face recognition (FR) share similar processes with FER, face recognition (FR) system was also explored. The methodology is focused on confirming the hypothesis that LDR images are limited in pixel representation on capturing the full scene light present in the harsh lighting environment, and in particular answering the question: *"can the same emotional face be recognised under different changing lighting conditions?"* These are presented in Chapters 5, 6 and 7.

The main objective of this research is to explore how FER tasks can be improved in a way that image data employed will be made insensitive to changing light and the existing methods used to boost FER tasks. FER systems are highly

dependent on the quality of images that are provided. For FER to be useful in the real-world, it must be capable of working with images captured in complex lighting conditions. Pre-processing has been previously used to allow FER to cope with such images, although, pre-processing is capable of changing the perception of the resultant images [BJ11], there is no guarantee that most of the appearance details needed for FER will be preserved. HDR imaging offers a potential solution to current limitations.

4.2 The Methodology

Outline of the overall methodology for this thesis is presented in Table 4.1.

Table 4.1: Overall Methodology Outline

1	Create a straightforward HDR database (with elements of harsh lighting)	
	i. Check traditional methods on HDR database - single exposure, optimal and pre-processed datasets. ii. Check traditional methods on LDR database - to obtain baseline scores. iii. If result of i not equal ii, then harsh lighting has an effect.	To check the effect of harsh lighting
	Do HDR method performs better?	
	iv. On tone mapped images with FER using LBP and SURF. v. On HDR directly with FER using LBP and SURF (if iv fails). vi. If v is as good as ii and greater than i, then TM are good enough.	To check HDR performance
2	Create second HDR database (enhanced version)	
	i. Repeat experiment 1 for Face recognition using LBP and SUFR ii. Repeat experiment 1 for Facial expressing recognition using SURF and deep CNN	To check FER/FR performance

The research methodology that will be undertaken is as follows:

1. Investigate whether emotion recognition can be extended to HDR tone mapped images and also study the benefit HDR imagery holds for FER with data collected under harsh lighting conditions.

2. Explore whether HDR can benefit FR systems following the outcome of the investigation in 1.

3. Carry out a further analysis to strengthen the results in 1 and 2 above in order

to understand how to carry out a fair analysis and comparison of the results based on a HDR database created with three different lighting conditions (back-light, left-light and overhead-light) and presented in an experiment to a number of participants.

The overall experimental pipeline of image enhancement methods using pre-processing techniques, FER and FR experiments carried out using pre-processed datasets and HDR datasets is presented in Figure 4.1.

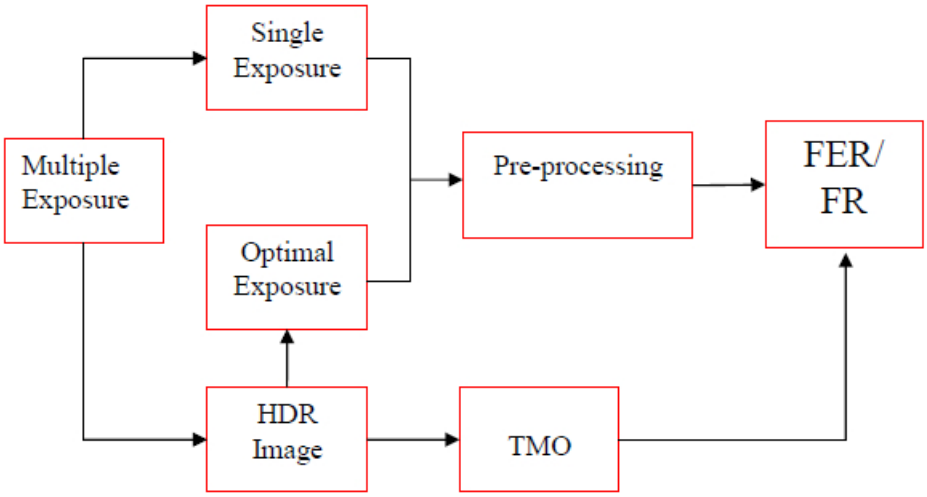

Figure 4.1: Overall Experimental Pipeline

The three experimental setup are discussed individually in the subsequent sections to show how these bring value to the overall objectives of the research.

4.2.1 Pre-processing/FER Research Method

Pre-processing is a common name for operations with images at the lowest level of abstraction aimed at improving image data that suppresses unwanted distortions or enhances some useful image features for further processing. For example, pixel brightness transformation, geometric transformation, etc. A pre-processing task was carried out in order to investigate whether, when images affected with harsh lighting conditions are pre-processed, there is a trade-off between loss in

visual quality and pixel distortions. In particular, we need to answer the question: *"Is pre-processing good enough and there is no significant loss of visual information"?* Previous work has shown that there is no guarantee of complete lighting invariance with existing methods [CBJ00]. Therefore, in order to establish a motivation for considering HDR imaging over the traditional LDR imaging, the following procedures were followed:

1. A set of images were captured using seven different facial expressions of emotions *(anger, disgust, fear, happiness, sadness, surprise and neutral)*. Scenes were created to allow images to be captured by placing light sources at 90^0 and 3 metres away from the subjects. Materials for the database image capture (database 1) are presented below:

    - Camera body: Canon EOS 5D Mark III.
    - Lens: Canon EF 24-105 f/4.0.
    - Shutter speed, from 1/1250 to 0.05 seconds
    - Bracketing range: -7 to +7 Ev
    - Number of exposures: 7
    - Bracketing interval: 1
    - Capture format: .cr2 (14 bits/pixel/channel)

    The set up for HDR-lab data capture is shown in Chapter 5.

2. The images were then pre-processed and, used to create two datasets: original images (orig-img) and pre-processed images (prepro-img). Pre-processing was achieved with four adopted image enhancements methods: adaptive histogram equalisation (AHE), normalised discrete cosine transform (NDCT), homomorphic filtering (HF) and gamma intensity corrections (GIC). These are described in Chapter 5.

3. In order to compare the original images against the pre-processed images, objective measures were adopted to quantify the pixel changes/loss and to determine which quality measure is important and which leads to some drawbacks. Three image quality measures (based on pixel-by-pixel distance measures), were chosen: entropy, structural similarity index (SSIM) and mean square error (MSE).

4. The usefulness of the pre-processing task was investigated with the two adopted public databases: Japanese female facial expression (JAFFE) and static facial expression in the wild (SFEW). The JAFFE database was transformed into mJAFFE - by creating artificial harsh lighting conditions.

5. Having established the benchmark, experiments were carried out to ascertain how FER can be improved. The facial regions from the databases were detected and resized by scaling to 75×75 pixels in order to extract the important parts of the face (in computer graphics, scaling does not change the image pixels or the image shape in any way). The scaled images were divided into 80% training and 20% validation. The faces were then used to compute LBP and SURF features and SVM was used for classification. The combination of LBP+SVM and SURF+SVM algorithms were implemented for FER in Matlab.

6. Validation of the HDR-lab database, mJAFFE and SFEW datasets was conducted in a subject independent manner. Afterwards, FER accuracy was presented on the pre-processed datasets AHE, WF, NDCT, GC and Naive (which were created from the HDR-lab), mJAFFE and SFEW databases.

4.2.2 Face Recognition Research Method

In line with ascertaining how FER can be improved, face recognition (FR) was also conducted for the purpose of creating a similar benchmark. As with FER, FR

performance can also drop under difficult lighting conditions in the applications to which it is applied. Low performance under uncontrolled lighting conditions can be due to expression, pose, gender, image resolution, illumination, etc. The objective of this was to establish whether HDR can benefit FR systems with the minimum possible intrusion. To achieve this, the following procedures were followed:

1. Based on the notion that different qualities of light can make our faces look very different, an extended HDR database (database 2) of people of different ages and ethnicities was created under three harsh lighting scenarios: one with a bright light source behind the face, one to the side producing harsh shadows across the face and a similar one at an angle producing shadows at an alternative angle. The database contains four emotions - angry, disgust, happiness and neutral. Other versions of the HDR database were created using the logarithmic and display adaptive TMOs. Naive (zero exposure), NDCT (pre-processed naive) and optimal exposure are additional datasets. Normalised discrete cosine transform (NDCT) has been described in Chapter 2, the optimal exposure dataset is an HDR to LDR method that modifies the presented HDR data into LDR by selecting the largest contiguous areas in luminance space to fit into LDR [DBRS*15a].

   The schematic diagram of the illumination setup is shown in Chapter 6, Figure 6.1, while the materials for the database 2 image capture is presented below:

   - Camera body: Canon EOS 5D Mark III
   - Lens: Canon EF 24-105 f/4.0
   - ISO: 1600
   - Aperture: $f/5.6$ (base exposure)
   - Bracketing range: -8 to + 8 Ev

- Number of exposures: 7
- Bracketing interval: 2.42 Ev
- Capture format: .cr2 (14 bits/pixel/channel)

2. The neutral emotions in the database were collected to create a neutral dataset for the FR task. The dataset was then divided into 80% training and 20% testing containing the same people. The SURF algorithm using BOF technique was used for feature extraction, while multi-class SVMs were trained to classify the image data into similar faces. To prevent data over fitting, the k-fold cross-validation was used. The setup was repeated five times and recognition rates averaged over the five trials.

3. The different experimental conditions were meant to validate the performance of HDR on the FR algorithm used. Therefore, results are presented across the three different induced lighting representations:

   - Summary of FR rates based on naive, NDCT, optimal exposure, logarithmic and display adaptive TMOs.
   - Summary of FR rates based on the three separate lighting conditions.
   - Performance of FR based on the naive (0^{th} exposure), NDCT, optimal exposure, logarithmic and display adaptive TMOs.
   - Finally, a summary of precision across the three lighting conditions and five datasets.

4.2.3 HDR Database Evaluation Method

One of the major challenges which remains unresolved with computer recognition systems is how to deal with changes in facial appearance (cue of the shape of facial features, like the eyes, nose, mouth, etc) due to harsh lighting conditions. To put this in perspective, further analysis of the two methodologies presented in Sections

4.2.1 and 4.2.2 is conducted in order to strengthen the findings on FER experiments. This is carried out as described below:

1. Using the extended database created in Section 4.2.2, the database, which contains four emotions was divided as per 80% training and 20% testing.

2. FER was conducted using two algorithms:

    - Feature extraction was achieved using SURF and bag of features and multi-class SVMs were used to train the classifier.
    - In addition, a convolutional neural network (CNN) was trained to solve the FER problem.

3. The purpose of this experiment is validating the level of performance that can be achieved with HDR database under different lighting conditions for FER tasks and to find out whether facial lights from different directions does have significant effect on facial recognition. Based on this, FER performance across the different induced lights under: all lights combination and all lights separated was adopted for testing the effects of changing light on facial expressions. Validation was done as follows:

    - Validating overall FER performance on the HDR database across 0^{th} exposure (naive), optimal exposure, display adaptive TMO, drago TMO, logarithmic TMO and reinhard TMO under the type based lighting conditions mentioned above. This is sub divided into:

    - Due to the nature of the scene light, FER with all lights combinations was conducted under the basic emotions (angry, disgust, happiness and neural) using naive, optimal exposure, display adaptive TMO drago TMO, logarithmic TMO and reinhard TMO and reported on the two algorithm used SURF+BOF+SVM and CNN.

- FER performance with all lights separated was conducted. The reason for this experiment is to enable a test of each light condition under the four emotions.
- In addition, in order to address the need to make our studies reliable for comparison in domains where computation of precision and recall are mostly useful, and to have a clear understanding of how close the classifier performance is between the three lights, precision and recall were computed.

4.3 Software and Programming Language

All experiments conducted in this research were achieved with Matlab. Matlab is a multi-paradigm numerical computing environment. It is a general purpose programming language, as a usage choice, it provides many functions for image processing and its related tasks, where most of the functions are written in Matlab language and are publicly available in readable plain text [Gil09].

4.3.1 Speeded-up robust features (SURF)

SURF algorithm was implemented with bag of features (BOF) approach using the following steps:

- Extract local features using SURF image descriptor from the images
- Put all local features into a single set (feature vectors)
- Quantise the feature vectors using clustering algorithm (k-means) to find centroids coordinates (vocabulary).
- Take quantised descriptors and create histogram (global feature vector)
- Create SVMs from histograms

- Perform classifications

4.3.2 Local Binary Pattern (LBP)

LBP algorithm implemented for FER and FR was achieved in the following ways:

- The algorithm was implemented on the database of faces.
- Face image are divided into local regions and LBP texture descriptors are extracted from each region independently.
- The occurrences of the LBP codes in the images were collected into a histogram.
- Use the LBP texture descriptors to build several local descriptors of the face and concatenate them into a global description of face.
- Classification was performed by computing the histogram similarities.

4.3.3 Deep Neural network (CNN)

Deep learning using deep neural networks (CNN) was implemented as follows:

- Train a classifier that can classify facial images into different emotions
- Use a model that has already been trained on a common type of images and adapt it to the problem.
- Setup MatConvNet
- Download ImageNet model from MatConvNet
- Train datasets of facial expressions
- Using transfer learning, perform feature extraction using pre-trained CNN for new tasks.

- Train a classifier using CNN features
- Perform emotion recognition.

Chapter 5

Facial Expression Recognition under Complex Lighting

"I have had my results for a long time, but I am yet to know how to arrive at them." Carl Friedrich Gauss

5.1 Introduction

Differences in facial appearance due to changing light can be a result of shadow, surface reflectance, specularity, occlusions, pose, etc. Non-uniform facial lighting conditions from different directions and intensities can result in inaccurate detection of features, thereby affecting the performance of recognition [HS07]. In the image processing research community, different techniques have been developed for either separating (extract) or suppressing (normalise) complex lighting conditions [TT10]. While these techniques have worked, they have limitations. For example, useful information can be lost or artefacts added after performing contrast stretching or illumination normalisation [VPG09] in the process of dealing with the over-exposure or the under-exposure of the image pixels. In [SGDM11], it was reported that the effects of harsh lighting are drastic

on pixel level, whereby the alteration of pixel information can falsely improve or reduce the performance of the applied algorithm. Therefore, a good classification algorithm for recognition should be invariant to the appearance differences, while retaining sensitivity to the inter-class variations.

In this chapter we hypothesise that HDR imaging methods can improve the task of FER for scenarios with complex lighting conditions. Frequently, improving image quality has relied on image enhancement techniques, where either contrast is stretched or illumination is extracted. With a traditional LDR pipeline the range of available brightness values is still rather limited. HDR techniqus provide the possibility of manipulating scenes with higher dynamic ranges [BADC11].

We use HDR imaging techniques to support FER without changes to the FER systems themselves enabling traditional methods to be used directly. This should allow for improved recognition of local features, which can align with the way that the Human Visual System processes emotions [BHL*10]. In particular, we adopted tone mapping operators which serve the purpose of compressing the luminance of HDR content to LDR.

5.2 Pre-Processing Techniques

An important aspect of FER research focuses on building systems invariant to lighting conditions. A number of these are enhancement methods, based on pre-processing techniques which attempt to reduce the issues related to lighting. The pipeline for image enhancement methods detailing the selected pre-processing techniques is presented in Figure 5.1.

Adaptive Histogram Equalisation

Adaptive histogram equalisation (AHE) [HYAB13] performs contrast limited adaptive histogram equalisation on an image. It operates on a small data region

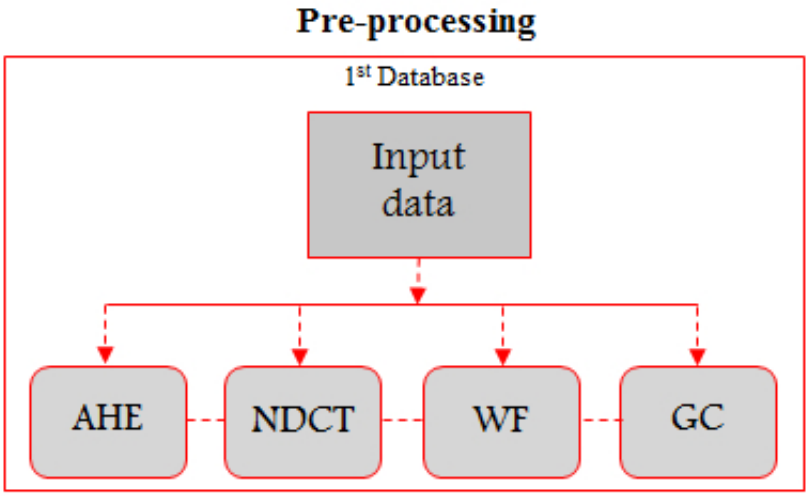

Figure 5.1: Pipeline for image enhancement techniques through different pre-processing steps.

(tile) rather than the entire image. Each tile's contrast is enhanced so that the histogram of each output region approximately matches with the specified histogram [VPG09]; that is, the uniform distribution by default, so that the contrast enhancement can be limited in order to avoid amplifying the artificially induced boundaries, which might be present in the image.

Normalised Discrete Cosine Transform

An illumination normalisation approach to reduce lighting variations, while keeping the main facial features unaffected has been adopted in [CEW06]. Taking into account that lighting variations lies in the low-frequency band [RdSQ08], the authors reported a significant reduction in lighting variations by discarding low-frequency DCT coefficients in the logarithmic DCT domain. However, shadowing and specularity problems were not completely solved in their work.

Wiener Filter

The Wiener Filter (WF) [GAKD00] operates as an inverse filter for noise smoothing. It performs deconvolution by inverse filtering (high-pass filtering), and removes noise with a compression operation (low-pass filtering). WF is derived from the power law spectra assumption of natural images and examines the statistical behaviour of face images under different lighting conditions [GAKD00]. In contrast to the existing methods [CYCC11] it assumes that facial illumination features lie in the high frequency spectrum. By analysing the power spectra of natural images, they show that some of these features could also lie in the low frequency part as well. They derive a Wiener filter approach to best separate the illumination-invariant features from an image.

Gamma Correction

The gamma curve (GC) function is often used to compensate for the nonlinear responses in imaging sensors or displays by raising the image pixels' intensity to a power called gamma. Recently, GC was used in [MG] to simulate the variance in light condition among images. Their test showed an effectiveness of the GC in different illumination conditions.

Another method [FQ14] applies a feature-based pre-processing method to generate a representation for light change invariance. This work performed a log-based transformation in part to reduce the effect of light changes. This leads to the expansion of the dark pixel values and the compression of the brighter pixels in the image by the log function, which generally results in a uniform spread of the pixel values. Furthermore, they transfer the scaled image to a Logarithmic fractal dimension image using the Differential Box-Counting algorithm.

The work in this chapter is on the effect of the facial appearance changes due to light changes when using HDR methods. As a result, issues with harsh lighting

are mitigated with HDR imaging, as dedicated space per bits enables the capturing of a HDR image than the eyes can perceive giving a realistic appearance. Some related studies [PMPP14] report automatic face (not emotional) recognition which was tested using sparse representations with tone mapped operators applied on HDR images. In [OPAHC*14a], they investigate the level of the performance that can be achieved for feature detection and tracking operations in the quality of images acquired with a HDR image sensor. Our approach follows by investigating the use of HDR methods for traditional FER systems.

5.3 Facial Expression Recognition Under Harsh Lighting Conditions

A robust FER system should possess the ability to classify the given image accurately irrespective of the lighting conditions. Techniques do exist to pre-process images to attempt to reduce lighting differences, such as histogram equalisation, discrete cosine transform (DCT), Gamma correction and logarithm transforms [SGCZ03, CEW06, STL*10]. However, a major issue lies in the ability of resolving the different emotions in environments with a high dynamic range such as those shown in Figure 5.2. As FER systems rely on computer vision techniques, they are highly dependent on the quality of the image that is provided. In complex lighting environments, with traditional imaging, parts of a scene may be over-exposed, while other parts may be under-exposed. HDR imaging ensures that all the detail in a scene is captured no matter what the lighting conditions.

In order to investigate the potential of HDR methods for FER systems, an HDR-based dataset of facial expressions under conditions with different lighting contrasts with people showing seven emotions was used. The dataset is used to evaluate various HDR methods for FER based on traditional techniques and compared with pre-processing methods. We considered two public facial

Figure 5.2: Sample images with harsh lighting.

expression databases, JAFFE [LAKG98] and SFEW [DGLG11a] for the purpose of performance validation. For feature extraction, we considered local-matching methods, LBP [SGM09] and an appearance based (pixel intensities) method, SURF [BETVG08a]. These techniques have been reported to be invariant to changing light in images [SGM09]. For learning, we use a multi-class SVM [HL02] that has been successful in image classification [FFP05a]. We present results for comparisons of unprocessed and pre-processed methods on all our databases and find that normalised DCT pre-processing performs best for the pre-processing methods. We subsequently compare the results with HDR tonemapping methods being used prior to the FER and find that tone mapping methods outperform the pre-processing methods significantly.

5.4 Facial Expression Recognition Method

Our overall method is based on LBP and SURF. LBP computes descriptors of a 256-bin histogram over a region of a texture descriptor. The derived binary numbers codify local primitives such as curved edges, spots and flat areas. The histogram of the labelled image $f(x,y)$ can be defined as:

$$H_i = \sum_{x,y} I(f_l(x,y) = i),\ i = 0,...,n-1 \tag{5.1}$$

where n is the maximum label number produced by the LBP operator and

$$I(A) = \begin{cases} 1, & \text{if A is true} \\ 0, & \text{otherwise.} \end{cases} \tag{5.2}$$

To extract features for recognition, the facial image is divided into 12 squares containing local information in 4 rows and 3 columns. On every pixel of each block, we apply the LBP operator, where we create a 59-bin histogram for each block. The histogram of the 12 blocks are then concatenated resulting in a single spatially enhanced histogram as a feature vector.

SURF extracts features using the [8 8] grid step and a blockwidth of [32 64 96 128], so that each image is a collection of detected patches over an 8×8 grid. Then a 500 bag of features (BOF) descriptors (or vocabulary) [FFP05a] is generated. This is then quantised using K-means to reduce the features [LM01]. For this experiment, after the fifth iteration, the recognition rate repeats itself. Therefore, the value of k was taken as 5. The number of centroids represent the number of features. The BOF provides an encoding method for counting the occurrences of the features in each image, where each image is assigned a membership to a large dictionary of the BOF and produces a histogram as a reduced representation of the given image into a feature vector. The encoded histogram as a feature vector is then used for training the multi-class SVM classifier.

In training the features, multi-class SVM [BETVG08a, SGM09] is adopted. The multi-emotion-class problem is solved by the one-against-all strategy to convert the six class problem into multiple binary classification problems in order to reduce over-fitting [VP12]. This strategy overcomes many real world challenges, such as the transitions between emotions [ZTC14a]. Each output of the binary classification is the confidence value of the test sample belonging to this class. The class label having the highest confidence is then used for final classification

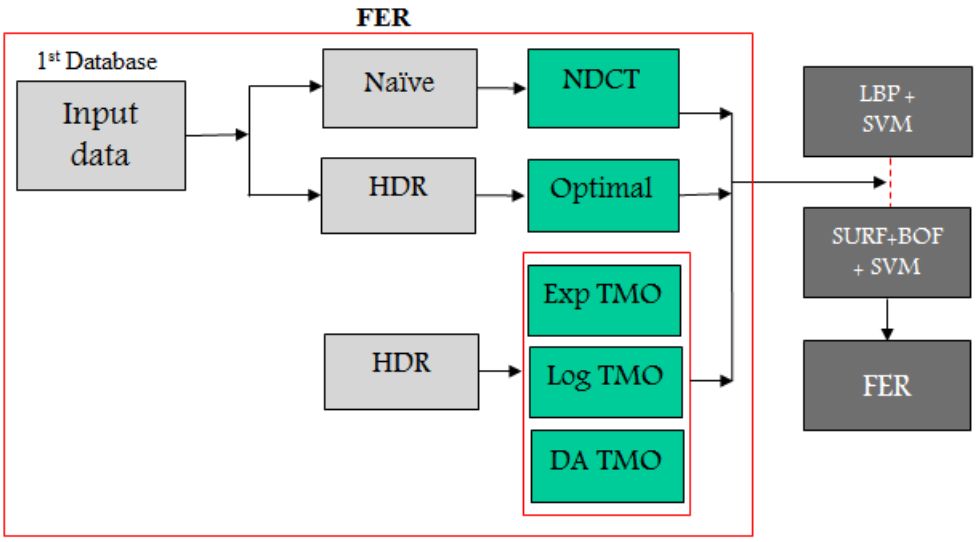

Figure 5.3: Facial Expression Recognition Pipeline.

of the emotion label for the corresponding image. We let $E_i(i = 1, 2, ..., N)$ be the emotion classes and $E_{i,j}$ be the emotion class produced by the SVM for the j^{th} image $(j = 1, 2, ..., M)$ in a training set C, where N and M are the total number of emotions and training set. The training set is classified as having the emotion E_i that has the majority voting among all the training sets:

$$C \in E_i \quad if E_i = max(\sum_{j=1}^{M} E_{i,j}). \tag{5.3}$$

. The facial expression recognition pipeline is presented in Figure 5.3.

5.5 Facial Expression Databases

In order to form an understanding of the impact HDR methods can have on FER systems in environments with harsher than normal lighting we used some existing databases for validation, modified one and created a new HDR database.

The HDR dataset:

The HDR dataset, which we shall refer to as HDR-lab, was created using seven

Figure 5.4: Sample of the 7 bracketed images from the HDR-lab dataset.

different facial expressions of emotions (*fear, anger, happiness, disgust, sadness, surprise and neutral*). Seven exposures were used to generate each HDR image (see Figure 5.4). Figure 5.5 shows the HDR captured scene. We created scenes to capture high and low contrasts, and dark and light shadows (vertical and horizontal) with the help of a light source placed at 90^0, 3 metres away from the participants.

To acquire the HDR images for our experiment, a Canon Mark EOS-5D III camera with shutter speeds from *1/1250 to 0.05* seconds was employed. In order to maintain control, an indoor scene was selected so that the conditions can be easily reproducible. As seen in Figure 5.4, the dynamic range of our images is large enough to require HDR.

The HDR-lab dataset contains 110 HDR images of ten subjects from different ethnic backgrounds and ages between 25-40. The average dynamic range of the images is 11 ($\log_2$). The dataset has some unique characteristics, in particular it is captured under harsh lighting conditions: shadows across the face, uniform bright light and dark areas.

JAFFE

The JAFFE [LAKG98] database has been extensively used for FER. It contains 213 images from 10 Japanese female actresses. It consists of six basic emotions (*anger, disgust, fear, happiness, sadness, surprise and neutral*). The mJAFFE dataset

Figure 5.5: Set up for HDR-lab data capture.

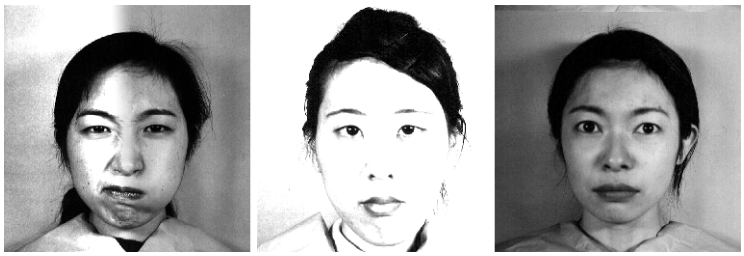

Figure 5.6: Sample images from the Improved JAFFE (mJAFFE) dataset

Figure 5.7: Sample images from the SFEW dataset.

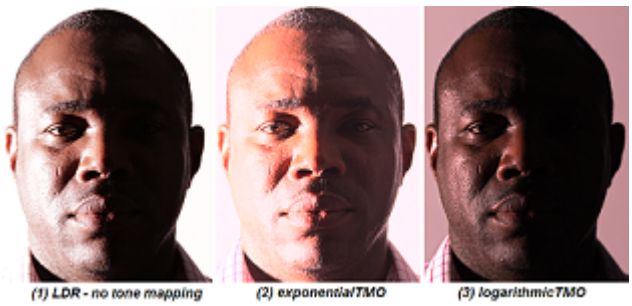

Figure 5.8: Original LDR and tone mapped images from the HDR-lab dataset.

was created using a Photoshop script to create shadows, high contrast, low contrast, over-exposed and under-exposed lighting conditions on the original JAFFE images. This is in order to impose harsh lighting conditions on the images. See Figure 5.6 for sample images.

SFEW

The SFEW [DGLG11a] dataset was extracted from frames from Acted Facial Expressions in the Wild (AFEW) database. The database was collected with close to real world lighting conditions, with different head poses, large age ranges, different face resolutions, occlusions and different focus. There are 95 subjects and a total of 663 well-labelled usable images. See Figure 5.7 for sample images.

5.6 Experiments, Results and Discussion

A number of experiments were conducted for validation with the JAFFE, mJAFFE and SFEW databases and to investigate the use of tone mapped HDR versus traditional methods on the HDR-lab database. Evaluation is conducted in a subject independent manner. We test for generalisation to new subjects using the leave-one-subject-out, k-fold cross validation [JVP11]. All images in the testing sets were excluded from training, so that no data from one subject appears in both the training and testing sets. For fairness, we randomly selected 10 subjects in each

dataset. For the subject independent method adopted the dataset was divided into n steps based on the number of subjects in the input data (in our case, n=10). Each step contains seven representations corresponding to the seven facial expression categories. Each representation is further divided into training and testing sets. The training set contains n *-1* subjects, and the test set contains n - *(n -1) = 1* subject. The division is carried out n times randomly and in a subject-independent manner, thus dividing the images on the basis of their facial expression labels.

The facial regions were detected using the Viola Jones algorithm [VJ01] and resized to 75×75 pixels in order to extract the important part of the face (according to [RR], resizing an image by scaling does not have effect on image pixel). The faces were then used to compute LBP and SURF features and SVM for classification. The LBP+SVM and SURF+SVM algorithms were implemented for FER in Matlab.

As a reference, our method gets scores of 70.8 and 70.5 for LBP+SVM and SURF+SVM respectively for the unmodified JAFFE database which is similar to results published in the literature for this database [DGLG11b].

In this work, we implemented four popular methods used for image enhancement: adaptive histogram equalisation (AHE) [HYAB13], Wiener Filter (WF) [GAKD00], Normalised Discrete Cosine Transform (NDCT) [CEW06] and Gamma Correction (GC) [Zha10] to serve as a comparison with the HDR methods. Initially we compare the results of these methods independently against each other. We also compare these results with unprocessed data (naive).

For HDR-lab we used Opt-exp as the method for deriving the LDR [NMP98, DBRS*15a]. Opt-exp selects the part of the luminance histogram that contains the most contiguous luminance content that can fit in 8 bits. Opt-exp is used to represent automatic exposure of a camera shooting in LDR.

Results showing the accuracy of the matchings are presented in Table 5.1. The NDCT approach outperformed the other approaches in all datasets. The FER accuracy on unprocessed datasets (naive) performed well below the other methods

indicating that pre-processing is useful and does improve results.

5.6.1 Results for the HDR methods

Based on the pre-processing results we compare the pre-processing Opt-exp (NDCT) method against the HDR tone mapping methods. We use different tone mapping operators - exponential (exp-tmo) [BADC11], logarithmic (log-tmo) [BQO*10], and Mantiuk et al's method (Mant-tmo) [MDK08a]. log-tmo and exp-tmo are straightforward operators that apply logarithmic and exponential functions to the HDR images respectively; they represent fast straightforward methods. Mant-tmo is a sophisticated tone mapper which performed very well against other tone mappers in comparison experiments [mel]. Sample images are shown in Figure 5.8.

Results are presented in Table 5.2. As can be seen the tone mapping methods all outperform the best of the pre-processing methods (NDCT) indicating that there is scope for traditional HDR methods to improve the performance of FER systems as the results begin to approach those of native methods on the more traditional databases such as JAFFE. In terms of the individual emotion classifications presented in Tables 5.6.1 and 5.6.1, all the tone mapping operators had less than 40% of mis-classified emotions. This is an encouraging performance.

In general, the performance based on SURF features achieved higher accuracy than the LBP features. It is important to note here that, our focus is on investigating whether emotion recognition can be extended to HDR tone mapped images and also to investigate how this can improve FER performance. From this regard, the two different methods were used to test for consistency across different aspects of FER systems and the HDR methods performed well with both.

Table 5.1: FER accuracy on pre-processed datasets (Opt-exp, mJAFFE and SFEW)(%). AHE=adaptive histogram equalisation, WF=wiener filter, NDCT=normalised discrete cosine transform and GC=gamma correction.

HDR-Lab					
Method	**AHE**	**WF**	**NDCT**	**GC**	**naive**
LBP+SVM	56.5	56	58.2	52	41.1
SURF+SVM	57.1	56.8	60.0	50.1	49.9
mJAFFE					
Method	**AHE**	**WF**	**NDCT**	**GC**	**naive**
LBP+SVM	59.5	55.7	70.8	62.2	27.8
SURF+SVM	61.8	49.8	65.1	65.1	42.7
SFEW					
Method	**AHE**	**WF**	**NDCT**	**GC**	**naive**
LBP+SVM	46.3	48.1	50.6	49	32.4
SURF+SVM	47.6	43.1	54.1	43.6	31.3

Table 5.2: Recognition accuracy with LBP+SVM and SURF+SVM algorithms (%) on the tone mapping operators for the HDR-lab dataset. Exponential (exp), Logarithmic (log), Mantiuk et. al. (Mant-tmo), N-Img=number of images.

Method	**N-Img**	**LBP + SVM**	**SURF+ SVM**
Opt-exp (NDCT)	110	58.2	60.0
Exp-tmo	110	66.5	73.5
Log-tmo	110	69.5	72.8
Mant-tmo	110	75	79.8

5.7 Summary

The traditional methods usually attempt to reduce the harsh lighting through illumination extraction or normalisation in the form of enhancement. Image enhancement is supposed to provide visual clarity and pixel understanding to the image processing algorithm [CEW06]. This raises the question: *"Does image enhancement improves image clarity and understanding without a trade-off between addition of artefacts and loss of information in the cause of this?"*

Table 5.3: 6-class classification accuracy with LBP+SVM (%).1st row of each dataset=correct classification, 2nd row of each dataset=mis-classification. ang=anger, dis=disgust, fea=fear, hap=happiness, sad=sadness, sur=surprise.

Emotion	**ang**	**dis**	**fea**	**hap**	**sad**	**sur**
JAFFE	71	69	70	70	68	77
mJAFFE	67	72	69	87	60	70
SFEW	50	52	51	55	32	64
Opt-exp	61	57	61	62	59	39
Exp-TMO	62	55	65	69	77	71
Log-TMO	62	82	51	67	86	69
DA-TMO	80	72	86	64	86	62

Table 5.4: 6-class classification rates with SURF+SVM (%). 1st row of each dataset=correct classification, 2nd row of each dataset=mis-classification. ang=anger, dis=disgust, fea=fear, hap=happiness, sad=sadness, sur=surprise.

Emotion	**ang**	**dis**	**fea**	**hap**	**sad**	**sur**
JAFFE	68	73	69	71	67	75
mJAFFE	56	67	66	56	58	68
SFEW	67	50	50	67	17	89
Opt-exp	58	56	68	63	63	35
Exp-TMO	81	59	68	72	80	81
Log-TMO	67	92	55	67	89	67
DA-TMO	89	76	93	67	89	65

In this chapter, a facial expression recognition (FER) study was conducted investigating whether HDR tone mapped images can improve FER performance under complex lighting conditions. To achieve this, a new straightforward facial expression dataset of HDR images, a collection of faces under different lighting contrasts described in Chapter 4 was used. Our approach applied a machine learning method, Support Vector Machines (SVMs) to texture-based, Local Binary Pattern (LBP) and appearance based, Speeded-Up Robust Feature (SURF) image representations to conduct emotion recognition experiments. We ran comparisons using HDR tone mapped images and more traditional pre-processing methods of reducing the impact of large lighting variations. The best of the adopted methods

achieved 75% and 79.8% accuracy which is significantly better than the more traditional pre-processing methods tested.

Chapter 6

Face Recognition under Complex Lighting Conditions

"Never discourage anyone who continually makes progress, not matter how slow." Plato

6.1 Introduction

This chapter explores whether HDR can benefit FR systems. In order to demonstrate the benefit, an HDR database of faces described in Chapter 4 was used. Using a traditional FR method for LDR data, FR rates were compared across a number of conditions. The main conditions considered in this chapter are LDR input for FR, LDR input pre-processed with traditional methods to reduce illumination issues, and finally HDR input converted to LDR using a traditional HDR optimal exposure and tone mapping method.

FR under harsh lighting conditions has received some attention in the research domain. The traditional approaches for dealing with this issue can be broadly described under normalisation based and feature based methods. In [TT10], the appearance based approach, part of the feature based methods,

was used to define training examples under different lighting conditions and directly used without undergoing lighting pre-processing and further used to learn possible illumination variations. These are then generalised to the variations present in the images. Also an approach involving modelling the effect of illumination on human faces was implemented in [LHK05]. They showed the existence of configurations of single source directions of light to be effective for FR and report that a linear subspace spanned by the corresponding images can be approximated to an illumination cone, thus performing better with FR under a wide range of difficult lighting conditions.

With the normalised based approach, images are reduced to a canonical form in a way to suppress the lighting variations [TT10]. For example in a face, naturally occurring incoming lighting distributions have information predominantly as low spatial frequency and soft edges, such that the high frequency information in the image are mainly signal (inherent facial appearance). The Multiscale Retinex (MSR) method of Jobson [JRW97] was used to cancel the low frequency information. The images were smoothed and used to divide the smooth version of itself. In a similar way, the Self Quotient Image (SQI) method [GB03] was used with a different local filter. This was improved [CYZ*06] with the Logarithmic Total variation (LTV) smoothing. Recently, a comparative analysis was conducted [SKM04] on the above and related methods. The authors concluded that the normalised based approach has been effective but limited in its ability to handle spatially non-uniform lighting variations.

The popular approach involving the extraction of insensitive illumination feature sets directly in a presented image, such as geometric features, image derivative, edge maps, Local Binary Patterns (LBP), local autocorrelation filter, Gabor wavelet, etc. [AHP06, PYL08, ZSCG07], perform well with raw grey images, but are limited in resistance to the harsh lighting variations present in real world scenes. Accordingly, the LBP features have proved to be effectively invariant to

monotonic global grey-level transformation [TT10], yet this method degrades in performance easily under harsh changes of lighting directions and shadows, and a similar performance drop also applies to other features discussed earlier.

Pereira, et al. [PMPP14] tested automatic face recognition using sparse feature representations with TMOs applied on 20 HDR faces from five selected individuals. They presented a preliminary result in which the logarithmic TMOs performed best. In this chapter, the work is conducted under controlled conditions with lighting environments clearly identified and quantified and with more participants. We carefully detailed procedures taken for the image capturing exercise. Furthermore, we generated more datasets for zeroth exposure and optimal exposure respectively from the original HDR database created, which contains 498 HDR images.

In order to obtain the HDR image, seven LDR images were bracketed within the range of R $\in [-8, +8]$ Ev with an interval of 2.42 Ev (exposure compensation value) between each LDR exposure. Since the exposure compensation values were changed, the sensor sensitivity/film speed (ISO setting) of the camera was fixed to 1600 and the base (0^{th}) exposure was fixed to a relative aperture of $f/5.6$. The shutter-speed was allowed to vary to obtain the required exposure. The capture details are given as below:

- Camera body: Canon EOS 5D Mark III
- Lens: Canon EF 24-105 f/4.0
- ISO: 1600
- Aperture: $f/5.6$ (base exposure)
- Bracketing range: -8 to + 8 Ev
- Number of exposures: 7
- Bracketing interval: 2.42 Ev

- Capture format: .cr2 (14 bits/pixel/channel)

The HDR dataset is composed of 21 participants with ages ranging between 23-50 years comprising of 17 males and 4 females. The average dynamic range of the images is 15 ($\log_2$). The LDR images were visually inspected and out of

Figure 6.1: Schematic diagram of the illumination setup.

focus/blurry (due to motion) images were discarded. Subsequently, the candidate images from each of the sessions were merged into HDR images, cropped at the facial region and scaled to a resolution of 150×150 pixels (as mentioned in Chapter 5, resizing by scaling does not have effect on image pixel). The resultant HDR images were stored in the .exr format [FKH] for further processing.

6.1.1 HDR tone-mapping

Face identification/recognition and feature detection algorithms such as Speeded Up Robust Features (SURF) and Scale-invariant Feature Transform (SIFT) typically operate on LDR images wherein the luminance pixel values accepted by the algorithms are in the normalised range of $\text{RGB}_{(x,y,z)} \in [0, 1]$. In order to take

advantage of traditional methods the captured HDR data is converted to the detection algorithm suitable format while keeping the overall appearance/tone of the scene as similar as possible to the reference HDR. Such a task is usually accomplished by either choosing one of the exposures (typically the base 0^{th} or optimal exposure [DBRS*15b]) or by using a TMO.

Analogous to the display driven data characteristics, the tone-mapped image is gamma corrected and passed to the feature detection algorithm. However, there are two major issues. Significant amount of details are unavailable in the base exposure and there is a plethora of TMOs which perform equally well for various requirements which presents a challenge of definitively choosing one. Since the goal in this work was maximal tone preservation, the Display Adaptive TMO proposed by Mantiuk et al. [MDK08b] was chosen as it has been classified as a Scene Reproduction Operator (SRO) [EWMU13] and has performed well in previous evaluations [MBDC15, UMM*10]. The results of another, more straightforward TMO, a logarithmic TMO as used by Perriera et al. [PMPP14] in their HDR study for facial recognition is also used as a comparison.

The goal of the SRO is to preserve the appearance of the original HDR scene including contrast, sharpness and colours by adjusting the image with the pre-notion of the ambient illumination and capabilities of the target display. The authors demonstrate that this can be defined as a non-linear optimisation problem which when simplified by reducing the degrees of freedom results in the introduction of a mapping technique with adjustable parameters. The SRO employs a piecewise linear tone-curve to map the HDR luminance to its corresponding LDR luminance. Given a particular display's characteristics, the TMO produces the least distorted image in terms of visible contrast distortions (measured in Just Noticeable Difference (JND) steps) which when weighed by an HVS model accounts for luminance masking, spatial contrast sensitivity and contrast masking. Moreover, the SRO also uses chroma preservation techniques

introduced by Schlick [Sch95] to preserve accurate chroma information. This SRO was chosen over many other candidates because it reproduces the reference HDR scene with minimal visible distortions and, as mentioned above, has performed well in previous evaluations. Figure 6.2 shows results of tone mapping on images in the captured database.

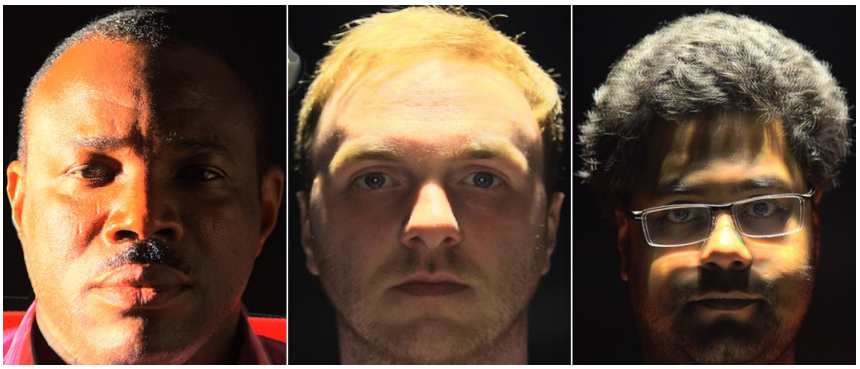

Figure 6.2: Sample tone mapped faces from our HDR dataset captured under different light.

6.2 Face Recognition

This section describes the adopted FR approach. Furthermore, the chosen method for illumination pre-processing for LDR is presented.

6.2.1 SURF using BOF Technique

The SURF algorithm using BOF techniques is discussed in this section as the next step following the image processing stage.

SURF

SURF (Speeded-Up Robust Features) is an invariant detector and descriptor using image interest points [BETVG08b]. The most important SURF property is the ability to repeat interest points, which goes to define how reliable and robust the

detector is in finding the same interest points under different lighting conditions, as in our case. Thus, the interest points are detected at distinctive image locations, such as corners, blobs and T-junctions.

Given a single channel image, I_0 with pixel intensity $I_0\ (x; y)$ at a given point $X = (x; y)$, if the central difference method is applied, the first-order difference achieved no response when applied to unchanging signals. To check for any changes in intensity, the second derivative of I_0(x; y) is expressed as:

$$\frac{\delta^2 I_0(x,y)}{\delta x^2} = I_0(x+1,y) - 2I_0(x,y) + I_0(x-1,y) \tag{6.1}$$

If the case that I_0 corresponds to another image I_u is considered, with an unknown lighting condition, and if we assume that the diagonal-offset models the two images (I_0 and I_u) as related by a linear transformation determined by a scalar constant α and an offset β, the pixel intensity can simply be modelled such that $I_u(x; y)$ of the image I_u at the same point $X = (x; y)$ is given as:

$$I_u(x,y) = \alpha I_0(x,y) + \beta \tag{6.2}$$

From Equation 6.2, the second derivative of $I_u(x, y)$ with respect to x, can be expressed as:

$$\frac{\delta^2 I_u(x,y)}{\delta x^2} = \frac{\alpha \delta^2 I_0(x,y)}{\delta x^2} \tag{6.3}$$

Similarly, the same principle can be applied to the second derivatives in y and XY, where the offset term β is cancelled out in the computation of the derivatives, without effect on the final result. But when the illumination is varied with a scalar α, there will be a proportional variation in the second derivatives with the scalar.

Typically, localising a feature with the SURF algorithm involves: interpolations, discarding low-contrast key points below a given threshold, (e.g.

threshold can be fixed at 0.03) and eliminating edges to increase key point stability. The challenge is, if a detector suffers from varying response to changing light, a feature (key-point) for instance in a bright image region may not be detected in the corresponding image with lower lighting levels. Therefore, the SURF detector response $R_u(x;y)$ of a given pixel $I_u(x;y)$ is given by the determinant of the Hessian matrix:

$$R_u(x,y) = \frac{\delta^2 I_u(x,y)}{\delta x^2}\frac{\delta^2 I_u(x,y)}{\delta y^2} - (\frac{\delta^2 I_u(x,y)}{\delta xy})^2 \tag{6.4}$$

Therefore, in order to compute SURF response, we substitute 6.3 into 6.4, so that the filter response R_u can be represented as $R_o(x;y)$ of the I_o and summarised into 6.5

$$R_u(x,y) = \alpha\frac{\delta^2 I_0(x,y)}{\delta x^2}\alpha\frac{\delta^2 I_0(x,y)}{\delta y^2} - (\alpha\frac{\delta^2 I_0(x,y)}{\delta xy})^2 = \alpha^2 R_0 \tag{6.5}$$

In **??**, because of the degree of α^2, even with small variation in scene light, there can be significant variations in the size of the detector response.

BOF

A Bag Of Feature (BOF) [FFP05b] algorithm is used in the study to build a feature dictionary. Like most modern approaches to category-level object detection, this algorithm attempts to use intensity features as input in both training and testing tasks. In this case, each image is represented as a collection of detected patches over a facial patch of 8×8 grid used to generate 500 visual BOF, where images are characterized by their illumination-invariant regions, along with their SURF descriptors. Quantisation is carried out through clustering to construct the image's signature formed by the centres of the clusters and their relative sizes into a more manageable size. The centroids is then used to provide an encoding method for counting feature frequency in each image, where each image membership is used to build a histogram of length k where the $i'th$ value is the frequency of the $i'th$ dictionary feature. Multi-class SVM is used to train on the object's categories and

used for image classification.

6.2.2 Face Classification

SVMs (Support Vector Machines) are a widely used supervised learning algorithm for data analysis and pattern recognition [ZTC14b]. SVM implementation is adopted for this problem. A RBF (Radial Basis Function) kernel was used to map the inputs to a higher dimensional space. This has been successfully used previously for face recognition [GJ10]. For instance, given a pair of faces, the SVM is trained to determine whether faces belong to the same or different subjects, a sample image is shown in Figure 1.8. The SVM was trained using the encoded histogram represented in the form of feature vector from the 21 subjects of the HDR-tone mapped dataset using 80% randomly generated pairs and testing with remaining 20% randomly generated pairs. As previously mentioned, we implement three induced harsh lighting conditions in our dataset to create three different representations, see Figure 6.1. Based on this, we decided to implement disjoint lighting conditions for training and testing. This means, we presented different input to the SVM for training and different input for testing.

6.2.3 Normalised Discrete Cosine Transform (NDCT)

One of the methods explored in the results section employs the use of pre-processing of the LDR image to remove aspects of the harsh lighting. The chosen method for pre-processing adopted for this work is normalised discrete cosine transform (NDCT). The authors in [IDMC16] compared four popular pre-processing methods for solving illumination problems used for image enhancement for emotion recognition, which shares similar characteristics to FR. It was found that the NDCT approach performed well above the other methods for the pre-processing methods. Also in [GNV11], the authors report lowest error rate with DCT even with down sampling coefficients.

The NDCT transform image representation as a sum of sinusoids of varying magnitudes and frequencies. Most salient information exists in low frequency coefficients. Unfortunately sometimes, the region where illumination varies are present in coefficients with low frequencies. Typically, illumination variations are reduced by setting the low-frequency DCT coefficients in logarithmic domain to zero [FN09b]. In the frequency domain, it is widely believed that illumination changes mainly in the low-frequency band, and research [GNV11] has shown that illumination changes slowly in the facial region. This means that, in order to obtain robust facial features under harsh light, the recovery of reflectance characteristics is important. Therefore, for illumination compensation, setting the DCT coefficient to zero is equivalent to subtracting the DCT basis image product and the corresponding coefficients from the original image. Setting the n low frequency DCT coefficients to zero gives:

$$F^{'}(x,y) = \sum_{\mu=0}^{M-1}\sum_{\nu=0}^{N-1} E(\mu,\nu) - \sum_{i=1}^{n} E(\mu_i,\nu_i)\ i\ = F(x,y) - \sum_{i=1}^{n} E(\mu_i - \nu_i) \tag{6.6}$$

where $E(\mu,\nu) = \alpha(\mu)\alpha(\nu)C(\mu,\nu)Cos[\frac{\pi(2x+1)\mu}{2M}]cos[\frac{\pi(2y+1)\nu}{2N}]$

And the illumination component term can be regarded as $\sum_{i=1}^{n} E(\mu_i,\nu_i)$.

The pipeline for face recognition is presented in Figure 6.3.

6.3 Results and Discussion

This section presents a series of results. The different experimental conditions are meant to validate the performance of HDR on the FR algorithm used. To demonstrate this, results are presented across the three different induced lighting representations (back, left and overhead) both individually and as a whole. For these results, the accuracy of the algorithm in learning a set of faces from training images and then correctly recognising the same people from a test set of different

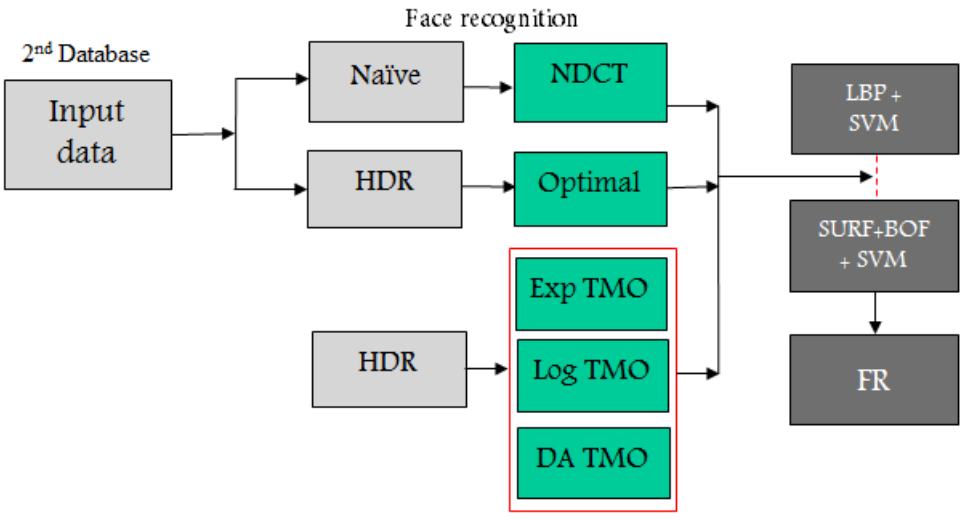

Figure 6.3: Face recognition pipeline.

images is evaluated. For these tests, both the training and testing sets contain the same people. To achieve this, tests are carried out based on 80% training and the rest on testing. This setup was repeated five times and the recognition rates averaged over the five trials. Due to limitations in the data set size a procedure known as cross-validation was adopted. This separates the data set ($N = 21$) into two parts leaving one part to be unknown, such that the prediction accuracy obtained from the "unknown" set more precisely reflects the performance on the classification of an independent data set.

To prevent data over-fitting, the k-fold cross-validation was used [ZTC14b], whereby the training set is divided into k subsets of equal size. Thus, one subset is tested using the classifier trained on the remaining $k - 1$ subsets. Where each instance of the whole training set is predicted once, consequently the cross-validation accuracy is the percentage of data that is correctly classified. Our classification problem falls into the category of the mutually exclusive (multivalue classification). In this case, a decision on one class leaves all options open for the other classes. Thus, in a sense the classes are supposed to be independent of one another, but the classes are rarely statistically independent. So multi-class SVM

classifiers are learnt and applied on each training set. Finally, the decisions of all classifiers is set as the recognition rate.

Initially the proposed FR method is validated against a traditional database. The Caltech 101 database [FFFP07], the de facto validation dataset for object recognition, containing images with large variations in light, pose and expression, was chosen for this test. In the comparative evaluation report of several recognition algorithms on Caltech 101 dataset conducted by [ZBMM06], the authors reported 65%, although their study was not limited to faces only. In this test a performance of 86% was achieved for the method presented in Section 8.3; this result was limited to testing only images that were clearly faces.

6.3.1 Overall FR performance

FR on the HDR dataset is tested across five methods: $0th$ exposure (naive), $0th$ exposure with NDCT pre-processed (NDCT), optimal exposure (optimal), logarithmic tone mapping (Lg_TMO) and Display Adaptive tone mapping (DA_TMO) with the type based lighting conditions - back light, left light and overhead light. The traditional approach is based on the set consisting of the $0th$ exposure from the seven exposures captured within bracketing range (-8 to $+8$). Since the naive approach is unable to capture the full scene lighting for the HDR scenarios, NDCT described in Section 6.2.3 was adopted and used as a pre-processing technique. This method would represent the LDR approach of dealing with harsh lighting. As in the previous chapter, the optimal method is an HDR to LDR method that modifies the HDR data into an LDR by selecting the largest contiguous area in luminance space to fit into an LDR. Lg_TMO is a logarithmic tone mapper that is considered relatively straightforward compared to other tone mappers. It is added here to provide a comparison with results in the related work of Perriera et al. [PMPP14]. DA_TMO was based on the DA_TMO as discussed in Section 6.1.1 and Lg_TMO for comparing our result with the work in

[OPAHC*14b]. This represents the state of the art of HDR to LDR methods. Table 6.3.2 presents results for the four methods. The performance increases as expected naive (82%), NDCT (84%), Opt_exp (87%), Lg_TMO (87.7%) and DA_TMO (93%). Naive under-performs as expected and NDCT does not improve much. Optimal is better but the TMOs are best. Lg_TMO does not perform as well as DA_TMO as expected. DA_TMO does very well with an overall of 93% indicating that a robust TMO may be sufficient for general FR performance for scenarios with harsh lighting.

6.3.2 Comparison of FR performance with disjoint training and testing set

In this section, the performance when the datasets are separated across the three lighting conditions is presented.

Generally, with HDR imaging [MBDC15] there is the advantage of a capturing a wide range of available scene light. But when images are captured in a scene where the area of interest is away from the camera lens, the performance of such HDR imaging would be low. Therefore, we observe that using images captured under the left light for training or testing lead to a drop in performance as shown in Table 6.3.2, but slightly higher in Lg_TMO and AD_TMO. Tables 6.3.2, 6.3.2, 6.3.2, 6.3.2 and 6.3.2 show the confusion matrices for the recognition. The confusion matrix is used to display statistics for assessing supervised classification accuracy, all correct guesses are located in the diagonal of the table with degree of mis-classification among classes (errors) shown on the outside. As reported, DA_TMO performs best and is not producing many false positives.

So far, we have used the FER accuracy parameter to discuss the correctness of our classifier's performance. To further strengthen our discussion, we explore other clues to give a further understanding of where the classifier is going wrong. Since we adopted the multi-class classifier, for the purpose of generalisation, we

decided to compute the success rates of the presented confusion matrices using the *precision* and *recall* metrics from one label versus all other labels. Precision, gives all the predicted labels (for example, the class Bck_light). Similarly, Recall, gives all instances that should have a label Bck_light, meaning how many of these were correctly captured? [SL09].

To compute precision and recall, the following shall be defined:

- Our confusion matrix tables assumes three possible output labels: Bck_light, Lft_light and Ovh_light.
- The diagonals of the matrices contains the *True Positives (TP)* for each label.
- The sum of a column would be the total number of instances that should have label X_light
- The sum of a row would be total number of instances predicted as a particular label X_light
- Given the above, the precision of a label X_light is computed as: $precision = TP\_X\_light/(TotalPredicted\_X\_light)$
- The recall of a label X_light is computed as: $recall = TP\_X\_light/(TotalLabel\_X\_light)$

We notice that all the TP (accuracy) in Tables 6.3.2 - 6.3.2 computed come out to be recall metric. Based on this, we decided to ignore the computation for recall. To compute precision, we take all rows as the emotional labels being predicted and all columns as the predicted emotional labels and used the expression given below: $precision = TP\_X\_light/(TP\_X\_light + FP\_X\_light)$. The result of our computation is presented in Table 6.3.2.

Table 6.3.2 focuses on providing further understanding of the classifier's ability in predicting the labels correctly. We recorded high precision with DA_TMO

in all 3 lights, followed by optimal. These are expected, as was also reported in the confusion matrices above.

Table 6.1: Face Recognition rates with naive, NDCT, Opt_exp and TMO datasets (%).

Data instance	Recognition rate
Naive	82
NDCT	84
Opt_exp	87
Lg_TMO	88
DA_TMO	93

Table 6.2: Recognition rate base on individual lighting conditions. BL (back_light), LL (left_light), Ovh (overhead_light) (%).

	Bck_light	Lft_light	Ovh_light
Naive	75	82	90
NDCT	76	86	89
Opt_exp	78	91	91
Lg_TMO	89	88	86
DA_TMO	91	93	95

Table 6.3: FR based on Naive ($0^{th} exposure$). Recognition rate 82%.

Data Instance	Bck_light	Lft_light	Ovh_light
Bck_light	**75**	13	12
Lft_light	7	**82**	11
Ovh_light	4	6	**90**

Table 6.4: FR based on NDCT. Recognition rate 84%.

Data Instance	Bck_light	Lft_light	Ovh_light
Bck_light	**76**	6	18
Lft_light	10	**86**	4
Ovh_light	9	2	**89**

Table 6.5: FR based on optimal exposure. Recognition rate 87%.

Data Instance	Bck_light	Lft_light	Ovh_light
Bck_light	**78**	10	12
Lft_light	1	**91**	8
Ovh_light	3	6	**91**

Table 6.6: FR based on Lg_TMO. Recognition rate 87.7%.

Data Instance	Bck_light	Lft_light	Ovh_light
Bck_light	**89**	6	5
Lft_light	8	**88**	4
Ovh_light	8	6	**86**

Table 6.7: FR based on DA_TMO. Recognition rate 93%.

Data Instance	Bck_light	Lft_light	Ovh_light
Bck_light	**91**	0	9
Lft_light	7	**93**	0
Ovh_light	1	4	**95**

Table 6.8: Summary of Precision across the 3 lights and 5 datasets (%)

	Bck_light	**Lft_light**	**Ovh_light**
Naive	87	81	80
NDCT	80	92	80
Opt_exp	95	85	82
Lg_TMO	85	88	91
DA_TMO	92	96	91

6.4 Summary

This chapter has presented the use of HDR in face recognition. Results demonstrate that traditional methods struggle to recognise faces under the complex lighting conditions presented here. However, the use of straightforward HDR techniques, by just capturing in HDR and tone mapping provided good performance.

Crucially, this approach does not require the development of new FR systems but can be used on already functioning methods without modification.

Chapter 7

Evaluating FER under Harsh Lighting with an Enhanced HDR Database

"As for the future, it remains unwritten. Anything can happen, and often we are wrong. The best we can do with the future is prepare and savor the possibilities of what can be done in the present."
Todd Kashdan

7.1 Introduction

The focus of this thesis is in using images collected from scenes where sudden changes in environmental lighting conditions may be expected. As has been shown in Chapter 5 and Chapter 6, sometimes, images can be captured with dynamic lighting conditions and are expected to be used to recognise facial expressions of emotions. Here, a traditional image processing approach may not be sufficient to deal with the lighting effect on the image.

Facial expression of emotion is displayed when muscles beneath the facial skin move. This movement can convey social and emotional information between

humans, and according to some researchers, they are the primary means of non-verbal communications. When images are presented under conditions where unstable lighting are expected, studies have shown that the differences in lighting conditions contributes more to image differences than changes in facial features and even more to variations in emotions [MM]. Therefore in the development of facial expression classification systems in such environments more attention is needed to changes in lighting conditions. Many studies have adopted the popular feature engineering technique to address the problem ranging from pre-processing of the training and testing images to lighting normalisation, the removal of lighting or the equalisation of lighting effects. Recently, [MM] carried out an experiment to rectify the effects of image lighting in order to increase classification rates in harsh lighting conditions. They reported improvement in classification based on Fisherface and SIFT keypoint repeatability. This is contrary to the work in Chapter 6, which showed that, although pre-processing is useful, it does not improve recognition results.

Over the years, computer vision research has included the principle of classical machine learning techniques, where features are defined by hand. Often features of interest are extracted and learning models trained to recognise/classify the presented data. However, there is a simple way of doing this through training the Convolutional Neural Networks (CNN). The CNNs with Deep Learning model have been used to achieve state-of-the-art accuracy in object recognition [MCM16]. Deep Learning models are trained using a large set of labelled data and architectures containing many layers. This chapter presents a further analysis to strengthen the results of the previous two chapters.

7.2 Facial Expression Recognition under Harsh Lighting Conditions

Recent work in FER related to scenarios where most often dynamic lighting conditions can affect recognition accuracy have increasingly focused on improving either data acquisition techniques or the achievement of a robust algorithm through fine tuning. The similarity of their work to this, is the use of images affected with harsh lighting, this is similar to the enhanced database adopted in this chapter. In Chapter 6, an HDR technique was used to improve FER performance with an improvement of up-to 2% based on the performance comparison of two TMOs. Similar to this study is [ZTC14a], they investigated images of people showing different facial expressions from television broadcasts and the World Wide Web for FER performance; they introduced fully automatic systems using a fusion based approach. Improvement in performance was gained through the fusion of point-based texture and geometric features. In this study, HDR tone mapped images showing emotional faces under harsh lighting conditions is used for FER through the use of SURF, LBP and CNN. As the work in [ZTC14a], the focus is on improving the performance of FER under harsh lighting conditions. In addition, in Chapter 6, this study show that sufficient improvement can be achieved by adopting HDR tone mapped images to solve face recognition problems in order to improve the performance of traditional face recognition algorithms when presented with data from scenarios with harsh lighting conditions.

The FER approaches were based on the traditional computer vision technique which rely on handcrafted feature detection, feature extraction and classification model. Recently, Mallahosseini et al [MCM16] proposed the deep neural networks approach to address FER problems across multiple well-known standard face datasets. They classify the registered facial images into one of the six basic expressions or neutral expressions contained in the datasets. They show

that the results of the proposed deep neural network architecture is better than the state-of-the-art methods and also better than the traditional neural networks in both accuracy and training time gained.

Similarly, Liu et al [LLSC13], investigated FER problem using a proposed AU-Aware architecture. They used convolutional layers and max-pooling layers to generate a complete representation of the presented faces, while the receptive field layer is used to generate a complete representation over all the possible spatial regions by carrying out a greedy convolution of the dense-sampling facial patches with spatial filters. The multilayer Restricted Boltzmann Machine (RBM) is used for learning the hierarchical features; further, the network outputs are concatenated as features to train a linear Support Vector Machines (SVMs) to classify the six basic expressions. They reported competitive results superior to LBP, SIFT, HoG and Gabor on the CK+ and MMI databases.

FER techniques were presented in the work of Kahos et al [KPB*13] which classified emotions expressed by human subjects in a short video clips extracted from feature length movies, where multiple deep neural networks were combined for four data modalities. More related to our work is the analysis of facial expressions within video frames using a deep convolutional neural network is used. Based on this, the best single model for training the convolutional neural network to predict emotions from static frames using two large data sets are: the Toronto Face Database and face images from Google image search. This yielded a test set accuracy of 35.58%. Using this strategy, accuracy of 41.03% was recorded on the challenge test set. This compares favourably to the challenge baseline test set accuracy of 27.56%.

7.3 Choice of Facial Expressions

The standard for most studies involving FER of emotions is to use the six universally recognised facial expressions. These facial expressions of emotions are based on the prototypic expressions defined in the Facial Action Coding System (FACS) [CAE07] as the target. In the previous chapter, the faces used to analyse the effect of the changes in lighting conditions was the average response across six expressions. However, to most actors of facial expressions, some expressions can be more difficult to act and even more difficult to discriminate by some algorithms because of the ambiguity or complexity involved [Ado02]. Recognition of light and emotion conveyed by facial expressions was investigated in [YF14]. The authors demonstrate that different facial expressions have different recognition rates. Thus, most frequently, happy and neutral expressions were correctly identified, while sad and disgust are least frequently identified [FCY16]. A subset of the six expressions is commonly used in both psychology and computer vision studies [WPB11, MJW*13] to investigate the influence of facial expressions on how we evaluate other people, and there is no consensus on the right number of expressions to use.

If we want to recognise facial expressions under harsh/changing lighting conditions, it is not enough to rely on the widely used databases. We need to employ a method that can overcome the limitations of traditional methods. For this reason, we created a prototype HDR-based dataset of facial expression under conditions with strong shadow across the face. This was described in Chapter 5. This database was used to evaluate various HDR methods for FER. To further investigate the effect of lighting on emotional faces, we created an extended HDR database under three different categories of harsh lighting conditions: side light, back light and overhead light as described in Chapter 6. It is comprised of four expression categories : neural, happy, angry and disgust. These are created under

three lighting conditions (back-light, left-light and overhead-light) for the purpose of investigating the influence of changing light on facial expressions of emotions. We decided to limit the expressions to four based on our experience in previous studies where actors find it difficult to distinguish between angry and sadness, disgust and fear.

In order to carry out the study, in addition to naive and optimal datasets, we considered datasets from four tone mapping operators: display adaptive [MDK08a], drago [DMMC03], logarithmic [BLDC06] and reinhard [Rei02], as representatives of the popular tone mappers for rendering HDR images on conventional LDR displays. Sample tone mapped images are shown in Figure 7.1. To test the HDR methods for FER algorithms, we used SURF+BOF+SVM (SBS) model and deep CNNs (AlexNet network).

HDR methods with emotional faces in the presence of three lighting conditions are adopted. We investigate FER performance presenting emotional faces under combined lights. Based on the report of Fotios et al [FCY16], some facial expressions may be easier to detect than others under changing light. For that reason, we also investigate FER performance presenting emotional faces under separate lights.

7.4 High Dynamic Range Database

The HDR database contain faces showing facial expressions of emotions in the category of the prototypic expressions defined in the FACS. There are 21 subjects distributed across 17 males and 4 females displaying four facial expressions (neutral, anger, happiness and disgust) under three lighting directions (back light, left light and overhead light). The average dynamic range of each image is 15 in ($\log_2$). We calculate the dynamic range in order to know the illumination conditions in the dataset. Dynamic range is defined as $\log_{10}(L_{max}/L_{min})$, where

L_{max} and L_{min} are the maximum and minimum luminance HDR brightness values, respectively. A comprehensive overview of the HDR database is detailed in Chapter 6.

Figure 7.1: Sample tone mapped images from HDR database

7.4.1 Speeded-Up Robust Features, Bag of Features and Support Vector Machines

Speeded-Up Robust Features (SURF)

We combine SURF technique with bag of feature (BOF) method before using the SVMs for expressions classification. The SURF detector has a history of providing greater invariance to changing light [IDMC16]. SURF searches for keypoints over a subspace of $\{x, y, \sigma\} \in \Re^3$. Based on the nature of the HDR database, we use the 'grid vector' method adopted from [VGV*16] to extract feature descriptors (feature vectors), where the grid is defined by step size of 8×8 using a blockwidth of [32 64 96 128] as the image patch size to extract the SURF descriptor. The grid vector captures the information about spatial arrangement of the SURF feature points. This gives us the advantage of extracting more features in images that are not likely

to contain distinctive features.

Bag of Features (BOF)

The BOF process is an adaptive approach to model image structure in a robust way, thus allowing for the identification of visual patterns relevant to the whole image collection [CCG09]. The standard for BOF is to represent data items (images) as a histogram over features. This idea has been previously successfully applied to some object recognition problems using HDR images under complex lighting conditions. Our data is not an exception, since it has a peculiar structure such as edges and variety of texture information.

It is desirable to have a representation based on image features that are stable and possesses certain invariant properties. This makes the BOF a potentially appropriate representation for the kind of visual content needed. The algorithm iteratively group the descriptors into k mutually exclusive clusters using k-means. A descriptor is categorized into its cluster centroid using a Euclidean distance metric. This parameter provides the model with a balance between high bias (underfitting) and high variance (overfitting). The clusters resulting from this process are separated by similar characteristics in bunch form, where each cluster centre represents a feature.

For a query image, we map each extracted descriptor into its nearest cluster centroid. A generated histogram count (frequency) of visual features (vocabulary) is constructed by incrementing a cluster centroid's number of occupants each time a descriptor is placed into it. Finally, each image is represented by a histogram vector of length N.

7.4.2 Classification with Support Vector Machines (SVMs)

The learning task uses the SVMs algorithm to classify the image base on its BOF. We consider the non-linear kernel, such that χ^2 kernel is given by

$$K(x,y) = 1 - 2\sum_{i=1}^{n} \frac{((x_i - y_i)^2}{x_i + y_i} \tag{7.1}$$

The facial expressions data should be classified into four emotion categories: anger, disgust, happiness and neutral. Thus multi-class SVMs is adopted, as it has previously been reported to be a robust scheme [ZZL*15] with the one-versus-all classification system, where the SVMs is trained to either classify an image as belonging to class c or class $\neg c$. Different SVMs are trained, where the number of SVMs needed grows linearly with the number of classes N, such that for the training data:

$$\{(x_i, y_i)\}_i^m, y_i \in 1, ..., N \tag{7.2}$$

a given multi-class SVMs will be used to train N separate SVMs. Thus to optimize the dual optimization problem [SM11]:

$$\max_a W(\alpha) = \sum_{i=1}^{m} -\frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j K(x^{(i)}, x^{(j)}), \tag{7.3}$$

where $K(x,z)$ corresponds to one of the Kernel functions. A query image is then classified using:

$$sgn\{\sum_{i}^{m} \alpha_i y^{(i)} K(x^{(i)}, z)\} \tag{7.4}$$

where $sgn(x)$ is an operator that returns the $sign$ of its argument and z is the query vector of BOF counts. Since values in coordinate visual features are between zero and one, no further scaling process on their values is performed before feeding these vectors into the SVM.

7.4.3 Deep Learning Approach

The accuracy of a DL model depends on the amount of data used to train the model. The most accurate model using a Convolutional Neural Network (CNN) requires thousands/millions of data samples to learn the weights for a classification problem. This can take a long time to train, however, a typical alternative to training a CNN from scratch is to use a pre-trained model (transfer of learning) that uses an optimized GPU to extract features from a new data set automatically. Compared to writing new CNN components, this is an important simplification that can significantly accelerate the application of the DL model without the need for a huge data set and very long training time. Once a DL model is trained, it can be applied to many applications, therefore it is logical to extend DL techniques to our FER problem.

Deep Learning (DL) uses multiple nonlinear processing layers to learn useful feature representations directly from the data. In this chapter, CNN architecture of the DL model is used [MMMK03] directly on the presented image data instead of training a machine to perform image classification. DL provides good image understanding, particularly in learning features directly from images used for classification, thus reducing the need for manual feature extraction and offering the benefit of extracting undefined features from the training data.

A CNN [VL15] is a function f mapping data $\mathbf{x}$ (such as image), to an output vector $\mathbf{y}$. The function $f = f_L \circ ... \circ f_1$ is the composition of a sequence of simple functions f_l, known as the "computational blocks or layers". Let $x_1, x_2, ..., x_L$ be the outputs of each layer in the network and let $x_\circ = x$ denote the network input. Each output $x_i = f_i(x_{i-1}; w_i)$ is computed from the previous output x_{i-1} by applying the function f_i with parameters w_i. A spatial structure representing the feature fields of the data flowing through the network is denoted by a 3D array $x_i \in \Re^{H_i W_i D_i}$, where the first two dimensions of the array are spatial coordinates.

For checking the network efficiency, the existing fourth non-singleton dimension allows for parallel batch processing of images. Generally, the f_i functions make the networks convolutional because they are non-linear filters operating as local and translation invariant operators.

Training a Deep Learning Model

The HDR database contains 498 images for 4 emotional classes. This is insufficient for training a DL CNN model. To train a DL model for FER with the HDR database, we adopted the AlexNet pre-trained networks [DDS*09], one of the popular pre-trained CNNs trained on the ImageNet dataset with 1000 object categories and 1.2 million training images. It has been established that pre-trained CNNs on a large collection of different image data are able to generalise well on scenarios that the CNNs has not been trained on. As pointed out in [LLSZ15], the pre-trained CNNs is shown to outperform the manual feature extraction techniques using SURF, HOG (histogram of oriented gradient) and LBP (local binary patterns).

Classification with Deep Learning Model

Traditional neural networks are structured layers consisting of a set of interconnected nodes. A CNN convolves learned features with input data and uses 2D convolutional layers. This makes the architecture well suited to processing 2D data, such as images. As described earlier, CNN also doubles as a classifier. Since we are using a pre-trained CNNs, our classifier is described as follows [VL15]. If we let the output $\widetilde{y} = f(x)$ be a vector of probabilities, and taking one each for the 1,000 possible image labels (faces, horses, hat, etc). And if the label of our image x is y, then the loss function $L_y(\widetilde{y}) \in \Re$ is used as penalty to classification errors. Therefore, learning can further be carried out on the CNN parameters in order to minimise the average losses over the datasets.

7.5 Results and Discussion

In this section, a series of results are presented for experiments carried out to evaluate performance of FER on the enhanced HDR DB. Performance across different lights representations (back, left and overhead) combined lights and separate lights were evaluated for the purpose of testing the effect of harsh light on emotional faces.

For these results, the accuracy of the algorithm in learning a set of faces from training images and then correctly recognising the same people from a test set of different images is evaluated. We divided our data into two subsets for both techniques - SURF and CNN. The first subset is the training set, which is based on 80% of the data. The second subset is the validation set, which is based on 20% of the data. The training and validation sets contain the same people. To avoid bias/variance in our results, we use Monte Carlo [JL10] cross-validation to randomly repeat the iteration five times and the recognition rates averaged over the five trials [ZTC14a]. With the SURF technique, the classification problem falls into category of mutually exclusive one (multivalue classification). As discussed in the previous chapter; multi-class SVM classifiers are thus learnt and applied on each training set. Finally, the decisions of all classifiers for SURF and CNN is set as the recognition rate.

Dataset

We generated six datasets from our HDR DB, comprising of four TMOs (display adaptive and logarithmic, drago and reinhard), including zero exposure and optimal exposure datasets. For the FER task, facial alignment or adjustment may distort or reduce the expression feature, so we simply use the original images that are cropped with Photoshop.

7.5.1 Overall Face Recognition performance

FER is evaluated on the HDR DB across six methods - $0th$ exposure (naive), optimal exposure (Opt_exp), display adaptive tmo (DA_TMO), drago tmo (Dr_TMO), logarithmic tmo (Lg_TMO) and reinhard tmo (re_TMO) with the type base lighting conditions - back_light, left_light and overhead_light. Due to the nature of our scene light, we investigated the performance of FER with the naive to confirm our hypothesis that the naive exposure is unable to capture the full scene light for the HDR scenarios. Lg_TMO is a logarithmic tone mapper that is considered relatively straightforward compared to other tone mappers. It is added here to provide a comparisons with the work in Perriera et al. [PMPP14]. DA_TMO was for the purpose of comparing our result with the work in [OPAHC*14a].

7.5.2 Facial Expression Recognition with Combined Lights

In this section, reports of the tests carried out to investigate the performance of FER with combined lights (left light, side light and overhead light), under the basic emotions: angry, disgust, happiness and neutral are presented using six datasets: naive, Opt_exp, DA_TMO, Dr_TMO, Lg_TMO and Re_TMO. with SBS and CNN techniques described in Section 7.4.1. For each of the emotions, for example, angry, all angry faces are collected from the three categories of lights. The same pattern was followed for disgust, happiness and neutral faces respectively.

To measure the supervised classification accuracy, Table 7.1 and Table 7.2 presents the confusion matrices for the classified categories of emotions with the six datasets using SBS and CNN techniques. These are used to present the statistics used for assessing the degree of misclassification among the classes. The actual label (value) is represented by a column and the predicted label (value) is represented by a row.

Classification Accuracy Based on SBS and CNN

In Table 7.1, Re_TMO comes top with 37% average recognition rate, followed by Dr_TMO 35% and Opt_exp 35%. On emotions, happiness appear as the easiest emotion for classification with 56% (Re_TMO), 54% (Dr_TMO), 52% (Lg_TMO) and 48% (DA_TMO). In contrast, angry is the most difficult emotion to classify in the datasets. This agrees with [ZTC14a], where they used data from television and the world wide web. They report that angry is mostly mis-classified as happy. They argue that realistic angry is characterised by mouth opening, which is also common with happy.

In Table 7.2, with the use of CNN model, there is performance improvement over what was presented in Table 7.1. This is expected based on our review of related works in Section 7.2. Relating it to this research, it was observed that recognition performance drops with combined lights as against the separate lights. This is traced to the different qualities of light on the skin, for example, non-white skin can be alienating, while white skin tend to be luminous [PCB01]. Also, with the tone mapping operators, Re_TMO top the list with 57%, followed by Opt_exp (52%) and Lg_TMO (49%). On emotions, happiness top the list with 67% (Re_TMO), 54% (DA_TMO and Lg_TMO), followed by other emotions. Similarly, angry also appears as the most difficult emotion to classify, this confirms our finding in Table 7.1.

Also in Table 7.1, the difference between the values of the predicted and actual labels, except **di**sgust (55%) for Opt_exp ; **ha**piness under Dr_TMO, Lg_TMO and Re_TMO. Similarly, in Table 7.2 **di**sgust (74%) for Opt_exp and **ne**utral (58%) for Re_TMO, others are **ha**ppiness under Re_TMO, DA_TMO and Lg_TMO.

The deep CNN technique used for this experiment rely heavily on convolution, it operates based on each pixel intensities and weights, wherefore

there will be high diffusion of pixel information in the initial image with its initial pixel intensities [Det15]. This means, strong diffusion will exist where there are changes in pixel intensities. For the detector, all information in the surrounding areas will concentrate in a single space. Consequently, pixels with lower values will most likely flow into the centre pixel and accumulate there. Thus, the largest concentration will be where the largest differences exist between neighbouring pixels. This says a lot about the performance recorded in Table 7.1 when different lights are combined. Overall, the result achieved accuracy of 57% (Table 7.2) with separate lights. This compares favourably with the 39.13% of the challenge baseline presented in [NNVW15]. Lastly, to achieve significant gain with deep neural network such as CNN seen in other domains, then having bigger datasets to train the network from scratch is crucial.

Table 7.1: Confusion matrix for combine lights - FER with DA_TMO, Dr_TMO, Lg_TMO and Re_TMO. (SBS)

Data	naive				Opt_exp				DA_TMO				Dr_TMO				Lg_TMO				Re_TMO			
	an	di	ha	ne	an	di	ha	ne	an	di	ha	ne	an	di	ha	ne	an	di	ha	ne	an	di	ha	ne
an	**27**	31	23	19	**31**	31	15	23	**27**	38	11	24	**27**	28	19	26	**30**	32	16	22	**26**	27	11	36
di	30	**31**	27	12	16	**55**	23	5	35	**24**	25	16	32	**21**	27	29	32	**21**	34	13	28	**23**	26	23
ha	22	31	**30**	17	25	23	**29**	22	13	12	**48**	27	14	15	**54**	18	11	14	**52**	24	10	12	**56**	22
ne	22	23	27	**28**	23	25	23	**27**	3	12	25	**33**	13	15	30	**43**	18	17	36	**28**	19	11	28	**42**
ReRt	27	31	30	28	31	55	29	27	27	24	48	33	27	21	54	43	30	21	52	28	26	23	56	42
AvRe	29%				35%				33%				35%				33%				37%			

Table 7.2: Confusion matrix for combine lights - FER with DA_TMO, Dr_TMO, Lg_TMO and Re_TMO. (CNN)

Data	naive				opt_exp				DA_TMO				Dr_TMO				Lg_TMO				Re_TMO			
	an	di	ha	ne	an	di	ha	ne	an	di	ha	ne	an	di	ha	ne	an	di	ha	ne	an	di	ha	ne
an	**31**	29	24	16	**42**	14	15	29	**30**	29	13	31	**43**	30	12	16	**45**	26	9	19	**53**	24	7	15
di	25	**45**	17	12	20	**40**	21	19	34	**40**	22	13	32	**46**	11	12	39	**42**	11	8	32	**47**	10	11
ha	15	17	**40**	27	13	13	**40**	34	9	19	**54**	18	11	14	**49**	26	14	9	**54**	24	6	8	**67**	19
ne	17	18	30	**34**	13	15	32	**40**	18	16	26	**40**	12	17	26	**55**	10	9	26	**55**	12	3	27	**58**
ReRt	31	45	40	34	42	40	40	40	30	40	54	40	43	46	49	55	45	42	54	55	53	47	67	58
AvRe	38%				41%				41%				48%				49%				57%			

7.5.3 Facial Expression Recognition with Separate Lights

Tables 7.3 and 7.4 present the summary of confusion matrices for FER performance over the six datasets and using SBS and CNN techniques based on the four emotions.

Classification Accuracy Based on SBS and CNN

In this section, FER performance on the HDR database with separate lighting conditions is presented. The reason for this experiment is to enable a test under each lighting conditions with the four emotions for comparison with the combined lights in Section 7.5.2. Comparatively, the true positives and average FER recognition rates indicates increase of about double the results presented in Tables 7.3 and 7.4, with combined lights. This indicate some traces of complexity in performance with combined lights. This result corroborates that of Chapter 6. We observe that, there is no significant difference to explain gaps between the experiments performed under different lights, even over the datasets. However, across emotions, Lg_TMO and Re_TMO perform better on average recognition rates than naive, Opt_exp and other TMOs with both SBS and CNN techniques. Similarly, naive performs lowest with SBS technique, but relatively higher with CNN technique. Globally, CNN performed better than SBS across all datasets. This is because deep neural network learn to identify shapes and objects to define facial expressions. Thus, deep CNN visible layer uses matrix, this enable the network to understand the spatial proximity of the pixels, leading to more robust feature extraction [YCBL14, NNVW15].

Table 7.3: Summary of confusion matrix showing true positives and the average recognition rates of FER on separate lights with Naive, Opt_exp, DA_TMO, Dr_TMO, Lg_TMO and Re_TMO datasets using SBS technique (%).

Emotion/	angry				disgust				happy				neutral			
Lights	BL	LL	OL	**AvR**	BL	LL	OL	**AvR**	BL	LL	OL	**AvR**	BL	LL	OL	**AvRe**
Naive	68	80	72	**73**	60	68	84	**71**	83	70	77	**77**	75	69	81	**75**
Opt_exp	87	82	91	**87**	72	64	60	**65**	87	82	85	**85**	74	80	82	**79**
DA_TMO	64	64	71	**66**	80	82	91	**85**	76	89	82	**82**	90	88	84	**87**
Dr_TMO	82	89	89	**87**	78	80	76	**78**	80	85	76	**80**	86	90	92	**89**
Lg_TMO	80	82	85	**82**	80	87	89	**85**	87	89	91	**89**	88	90	84	**87**
Re_TMO	87	93	80	**87**	82	76	80	**79**	93	93	85	**90**	82	86	86	**85**

Table 7.4: Summary of confusion matrix showing true positives and the average recognition rates of FER on separate lights with Naive, Opt_exp, DA_TMO, Dr_TMO, Lg_TMO and Re_TMO datasets using CNN technique (%).

Emotion/	angry				disgust				happy				neutral			
Lights	BL	LL	OL	**AvR**	BL	LL	OL	**AvR**	BL	LL	OL	**AvR**	BL	LL	OL	**AvRe**
Naive	82	80	91	**84**	84	87	84	**85**	86	86	86	**86**	84	82	78	**81**
Opt_exp	74	82	89	**82**	72	88	80	**80**	87	78	85	**83**	88	88	84	**87**
DA_TMO	59	68	67	**65**	71	71	78	**74**	93	85	87	**88**	82	88	88	**86**
Dr_TMO	80	84	78	**81**	76	71	89	**79**	85	76	76	**79**	88	80	84	**84**
Lg_TMO	65	87	80	**77**	82	82	80	**82**	87	85	87	**86**	84	90	84	**86**
Re_TMO	78	89	82	**83**	85	87	82	**85**	80	82	80	**82**	81	88	72	**81**

Discussion

The proposed HDR imaging was evaluated on a FER system using six test datasets. The test images consists of three different harsh lighting conditions - left light, side light and overhead light. The performance of the HDR based FER system under harsh lighting conditions using the combination of all three lights resulted in an overall performance of 37% as shown in Table 7.1 with SURF+BOF+SVM method. It is observed that the low performance with the BOF technique used for characterising (representing and discriminating) SURF descriptors, is due to the complexity of the combined lights under harsh lighting in the process of describing the structural information of the BOF in an image. In discriminating the different emotions, happy was highly discriminated among the emotions across all datasets and lighting conditions. This is confirmed in the test with SURF+BOF+SVM method. Unlike in Table 7.2 with deep CNN, where the overall performance was 57%. This is above the baseline of 55.6% in [NNVW15] in this domain. Although, HDR tone mapped images (DA_TMO, Dr_TMO, Lg_TMO, Re_TMO) performed much better than Naive and Opt_exp, which is the purpose of this experiment. On another perspective, comparing the results with other methods used for FER in the literature, the highest performance of 57% accepted for the image conditions used with deep CNN, still needs more improvement.

Similarly, in Table 7.3, the average recognition rates based on datasets

separated across the different emotional facial lights, recorded little difference with discriminating angry faces across the lighting conditions and datasets with SBS method. It is observed that most participants find it difficult making angry face. Also, in Table 7.4, there is no significant difference in discriminating all the emotional faces across the lighting conditions and datasets with deep CNN method. Taking the average recognition rates across the different lights, the left light (LL) recorded the highest recognition rate. This follows from the observation discussed above in Table 7.2. Therefore, it can be concluded here that these results can be used as a benchmark for other studies, with more attention on the complexities when images under different harsh lights are used in separate datasets or when combined in a single dataset.

7.6 Summary

In this chapter, an HDR solution applied to FER problem was investigated using data with different harsh lighting each presenting specific challenging conditions. We have evaluated the HDR database in six different categories. The performance of the six categories were assessed using two computer vision algorithms: SURF and CNN. We demonstrated under different lighting conditions that the tone mapped versions of the HDR database gives a high recognition rates, which are higher than the naive and optimal exposure.

Chapter 8

Conclusions

"If I were again beginning my studies, I would follow the advice of Plato and start with Mathematics." Galileo Galilei

The thesis has investigated the performance of a facial expression system using images captured under harsh lighting conditions. We assume throughout the thesis that the aim is on addressing the deficiencies and limitations of the native imaging method (LDR), and how this can be improved by HDR methods. While some progress has been reported in this direction, in this thesis, the path to achieving accurate, informative, robust and real-time facial expression analysis has been presented, particularly for images captured under harsh lighting conditions.

The images captured under harsh lighting conditions considered in the thesis are known to affect the performance of FER systems negatively. This challenge has been a major research issue both from academics and a practical perspective. However, in reality, most of the images captured will not be in ideal studio like lighting environments, where the face of the subject is perfectly lit and stable to ensure flawless capture. Key to the viable application of the methods is the ability to consistently measure the same facial expressions over the full range of changing scene lights. The main goal of the research presented in this thesis was to improve the performance of FER system by taking advantage of the HDR imaging

technology.

The thesis presented first, image enhancement methods used as pre-processing techniques, investigating how much of pixel information is resulted. Chapter 5 showed an experiment that uncovered the loss of image information resulting from image enhancement. The chapter also proposed HDR-based method towards improving the performance of facial expression recognition in scenes where harsh lighting conditions are expected and the LDR imaging find difficulty capturing the full range of scene light in a single exposure. Chapter 6 introduced an experiment that showed how the use of HDR tone mapping operators use for face recognition is beneficial for harsh lighting conditions. Finally, Chapter 7 present HDR database as a solution applied to facial expression recognition problem with different harsh lighting each presenting specific challenging lighting directions. Furthermore, the thesis contributions is given and suggestions for future work from the findings are also suggested.

8.1 Effect of Image Enhancement

Chapter 5 presented selected methods which exploits the traditional pre-processing approaches used for dealing with images affected with harsh lighting conditions. Image enhancement algorithms provides a multitude of approaches for modifying images to achieve visually acceptable images/conversion of presented images into a condition suitable for further processing by machine or human. The choice of image enhancement technique depends on the approach used for solving lighting condition problems. Thus, during this process of image enhancement, one or more of the image attributes are modified. The amount of modification wherefore, images pixel values are distorted, have been the subject of investigation in the thesis.

Four popular methods, Adaptive Histogram Equalisation, Wiener Filter,

Normalised Discrete Cosine Transform and Gamma Correction were implemented, their choice was based on the suitability for contrast enhancement. Structural similarity metric (SSIM) and correlation coefficient (CorrCoe) were applied to the original and pre-processed images as image quality measures. A comparison was made over five different trials; 68% and 74% average score was achieved using the AR-face database. This means an average of 29% loss of information from the original image (100%). Similarly, on the static facial expression in the wild database, over five trials where a comparison was made; 55.2% and 63% average scores was achieved; with average loss of information 40.9%. One important finding from this was that pre-processing directly affects the image information resulting in pixel distortion.

In order to further strengthen the investigation, an artificial lighting condition was introduced into the JAFFE database (called mJAFFE) the image lighting was increased by 10% and 20% and decreased by 10% and 20%. Repeatability measure using SURF features was conducted on mJAFFE and AR-face databases using the original and pre-processed images. The average of SURF repeatability measure between the images are, SURF 39% and AR-face 54%. This shows that there is a high loss of information between the original and pre-processed images.

This was applied to FER on six emotions (angry, disgust, fear, happiness, sad and surprise) using two image processing methods, SURF and LBP. The average recognition rates are thus, mJAFFE (86%), AR-face (82%), SFEW (56%) respectively. Based on these, it is possible to conclude, therefore that image enhancement by pre-processing in order to remove affected lighting conditions lead to the addition of artefacts or loss of information.

8.2 Facial Expression Recognition

Chapters 5 and 7 presented a method of facial expression recognition and Chapter 6 presents face recognition, these exploits the HDR tone mapped images. In Chapter 5, NDCT was selected as the best of the pre-processing methods. A comparison on FER with SURF and LBP using NDCT and HDR tone mapped images was carried out. Three different tone mappers: exponential, logarithmic and display adaptive were implemented. The display adaptive tone mapping method shows better performance against other tone mappers in the comparison. The FER accuracy for the HDR tone mapping methods are: exponential (LBP 66.5%, SURF 73.5%), logarithmic (LBP 69.5%, SURF 72.8%) and display adaptive (LBP 75%, SURF 79.8%). Also the accuracy with the pre-processed NDCT method was achieved as (LBP 58.2%, SURF 66.5%). All the tone mapping methods outperform the best of the pre-processing method (NDCT). This means that HDR relatively maintain the promise of solving some fundamental issues with better prospects for FER tasks.

8.3 Face Recognition

Similarly, Chapter 6, present a face recognition experiment to explore whether HDR can benefit FR systems. FR was conducted with five datasets - naive, NDCT, optimal, tone mapping methods (logarithmic and display adaptive) in the presence of three different lighting conditions: back-light, left-light and overhead-light. Recognition accuracy are thus: naive (82%), NDCT (84%), optimal exposure (87%), logarithmic tmo (88%) and display adaptive tmo (93%). The conclusion drawn from the results above demonstrate that HDR methods are beneficial for facial expression within harsh lighting.

8.4 Facial Expression Recognition under Harsh Lighting with an Enhanced HDR Database

Finally, Chapter 7 uses the SURF algorithm and Deep Convolutional Neural Network (Deep CNN) model to perform FER. For both SURF and CNN, HDR solutions were investigated for the FER problem using data captured under three different lighting conditions. Optimal exposure was used for LDR and HDR tone mapping methods (display adaptive, drago, logarithmic and reinhard) respectively. Two parameters were used for validation accuracy: tone mapping operators and emotion type. With SURF, performance was achieved thus, reinhard (37%), drago (35%) and optimal exposure (35%) top the validation accuracy for tone mapping operators. While happiness appear as the easiest emotion classified as 56% (reinhard), 54% (drago), logarithmic (52%) and display adaptive (48%); angry is the hardest to classify in the datasets. With CNN, recognition performance was improved. For tone mapping operators, reinhard top with 57%, logarithmic 49% and optimal exposure (LDR) 52%. Happiness also top with reinhard (67%), drago (54%) and logarithmic (54%). Angry also appear as the most difficult emotion to classify, confirming the findings above.

Dealing with image lighting problem in FER can be arbitrarily complex. But in this thesis, we have shown that there is a simply way of addressing the complexity. The experiments has been devoted to understanding the different effects of lighting directions. Three lighting directions were used in this chapter. In real-world scenarios, light directions may be difficult to anticipate, therefore we believe that our experiments can be applied in one or more of the lighting directions.

8.5 Contributions

The thesis make a number of contributions to the FER research community:

- The proposed HDR-based imaging techniques enhance the performance of FER in scenes with harsher than normal lighting conditions, and where LDR images have difficulty capturing the full range of light in a single exposure. This work was published in [IDC16]

- A first version of an HDR-based dataset was created called HDR-Lab. It contains facial expressions under conditions with difficult lighting contrast with people showing seven different facial expressions of emotions.

- HDR has been shown to play a key role in FER systems when dealing with environments that exhibit complex lighting. An enhanced version was created of the HDR-Lab database of faces showing four different expressions under three different harsh lighting positions - back light, left light and overhead light.

- After carrying out a study to evaluate the HDR database with FER tasks under different harsh lighting, the performance of two algorithms was measured: SURF and Convolutional neural network when all lights are separated and when all lights are combined. We note that with CNN, when all lights are combined, there are strong light diffusions of the pixel information, particularly where there are changes in pixel intensities. And for the detector, all information in the surrounding area will be flushed in a single space. Conversely, when lights are separate, we observed that there are complexities affecting the performance. This work was publised in [IDMC16].

8.6 Limitations

The limitations of this thesis are the following:

- The thesis only considers static images to solve FER problems, however, video data for FER would be fitting, although the method developed is efficient.
- Only a small number of faces and images were used.
- There is no labelled databases extracted from HDR images. This was addressed with the HDR database created in this thesis.
- In order to support the application of FER in the real-world, there would be more complex illumination to deal with. However, the techniques used in this thesis is enough to obtain plausible features for robust FER results. Maybe more stronger techniques can be employed to improve performance.
- Whilst the work in this thesis has been implemented considering algorithm efficiency and software re-engineering, the implementation presented could require some improvement with better software design and image choices.
- The traditional technique for acquisition of HDR images could lead to artefacts in bracketing, or reduce recognisable details or increase sensor costs [KSZ*14]. A more perceptually motivated approach would be an alternative, however, the traditional method adopted in this work is efficient.

8.7 Future Work

There are other aspects that requires further investigation in the work presented in this thesis.

Although, it is possible to record a large amount of information about a commercial packaging or product concept by taking video of a face, a lot

more can be captured by embedding the emotional facial imaging into a wider online survey and analysing it in-depth. The advantage is the inherent accurately assessed emotional reaction to brands and marketing stimuli with behavioural and demographic information is huge. Therefore, the need to investigate how facial imaging could help to improve marketing, advertising, retail displays, etc.

As quickly changing illumination are broader in the real-world, future work should investigate FER under complex dynamic lighting for dynamic expressions. In the case of emotion recognition, especially peak emotion in video, perceiving dynamics is of great importance. Atkinson et al [ADGY04] reports that observers perceive emotion dynamically, and dynamic expressions are thought to be easier to recognise than static.

Another area that is worth investigating is security applications of FER. Security modules in expert systems do not detect people's expression, thus giving allowance to commit crimes. Since facial expression correlate with feelings [HYHH11], it is necessary to investigate how to develop a real time FER system with potential of reading human facial expressions to identify those showing that they can cause harm or about to commit crime.

8.8 Final Remarks

This work has focused on the advantages of HDR for handling harsh lighting in FER methods. LDR images captured under harsh lighting conditions suffer from a loss of information and this also applies when using pre-processing techniques which attempts to improve image quality. This work demonstrated performance improvement with HDR methods directly over pre-processing of LDR images, in order to boost the amount of information captured in the original images and crucially allow them to be used with traditional FER.

This thesis has demonstrated performance improvement with HDR

methods directly over pre-processing of LDR images, in order to boost the amount of information captured in the original images and crucially allow them to be used with traditional FER. There are broader, long term issues that facial imaging could help address, for example, *"What if we could directly assess 'feeling' in real-world scenes without intricate scales or complex biometrics of neuroscience equipment?"* Technology such as HDR imaging, and in the future, perhaps even more advanced imaging techniques offer a potential solution. It is hoped that this thesis has provided a foundation upon which such future work work can build.

There are broader, long term issues to the industry that facial imaging could help address, for example, "What if we could directly assess 'feeling' in real-world scenes without intricate scales or complex biometrics of neuroscience equipment?" Technology such as HDR imaging, and in the future, more advanced imaging technique offer up exactly that prospect when utilised cleverly.

Appendix A

Experiment Consent Form

Investigator: Emmanuel Ige

Department: WMG, University of Warwick

Research title: Facial Expression Recognition under Harsh Lighting using High Dynamic Range Imaging

Purpose: The experiment is a facial expression capturing exercise under harsh lighting conditions, either in the bright day sunlight or places with casting shadows and low visibility.

Procedure: Participants are required to make seven different facial expressions learnt from the website that will be provided to registered participants. Before the experiment, each participant is expected to practice how to make the different facial expressions as described on the website.

Confidentiality: The captured images will be treated in the strictest confidence as they will be kept in a secure location at the University of Warwick, for ethical

consideration purpose. Although, the results of the study may be published or presented at professional meetings, but your identity will not be revealed. Participation is anonymous, which means you will not be identified from your image, as the only identification to the image is the type of emotion displayed.

You are free to withdraw from the experiment at any time prior to the conclusion, this means, your decision is paramount in taking part in the experiment, which means, you do not have to be in the experiment if you do not want to.

I will be grateful if you can contact me for participation on e.o.ige@warwick.ac.uk.

Bibliography

[ADGY04] ATKINSON A. P., DITTRICH W. H., GEMMELL A. J., YOUNG A. W.: Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception 33*, 6 (2004), 717–746.

[Ado02] ADOLPHS R.: Recognizing emotion from facial expressions: psychological and neurological mechanisms. *Behavioral and cognitive neuroscience reviews 1*, 1 (2002), 21–62.

[AHP06] AHONEN T., HADID A., PIETIKAINEN M.: Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 28*, 12 (2006), 2037–2041.

[Aky12] AKYÜZ A. O.: High dynamic range imaging pipeline on the gpu. *Journal of Real-Time Image Processing 10*, 2 (2012), 273–287.

[AMU97] ADINI Y., MOSES Y., ULLMAN S.: Face recognition: The problem of compensating for changes in illumination direction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 19*, 7 (1997), 721–732.

[AT12] ALLEN E., TRIANTAPHILLIDOU S.: *The Manual of Photography and Digital Imaging*. CRC Press, 2012.

[BADC] BANTERLE F., ARTUSI A., DEBATTISTA K., CHALMERS A.: High dynamic range imaging.

[BADC11] BANTERLE F., ARTUSI A., DEBATTISTA K., CHALMERS A.: *Advanced high dynamic range imaging: theory and practice*. CRC Press, 2011.

[BETVG08a] BAY H., ESS A., TUYTELAARS T., VAN GOOL L.: Speeded-up robust features (surf). *Computer vision and image understanding 110*, 3 (2008), 346–359.

[BETVG08b] BAY H., ESS A., TUYTELAARS T., VAN GOOL L.: Speeded-up robust features (surf). *Computer vision and image understanding 110*, 3 (2008), 346–359.

[BHL*10] BAL E., HARDEN E., LAMB D., VAN HECKE A. V., DENVER J. W., PORGES S. W.: Emotion recognition in children with autism spectrum disorders: Relations to eye gaze and autonomic state. *Journal of autism and developmental disorders 40*, 3 (2010), 358–370.

[BJ11] BASRI R., JACOBS D.: Illumination modeling for face recognition. In *Handbook of Face Recognition*. Springer, 2011, pp. 169–195.

[BKW*08] BREITENSTEIN M. D., KUETTEL D., WEISE T., VAN GOOL L., PFISTER H.: Real-time face pose estimation from single range images. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (2008), IEEE, pp. 1–8.

[BLDC06] BANTERLE F., LEDDA P., DEBATTISTA K., CHALMERS A.: Inverse tone mapping. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia* (2006), ACM, pp. 349–356.

[BLF*06] BARTLETT M. S., LITTLEWORT G. C., FRANK M. G., LAINSCSEK C., FASEL I. R., MOVELLAN J. R.: Automatic recognition of facial actions in spontaneous expressions. *Journal of multimedia 1*, 6 (2006), 22–35.

[BQO*10] BANDOH Y., QIU G., OKUDA M., DALY S., AACH T., AU O. C.: Recent advances in high dynamic range imaging technology. In *Image Processing (ICIP), 2010 17th IEEE International Conference on* (2010), IEEE, pp. 3125–3128.

[BS14] BRO R., SMILDE A. K.: Principal component analysis. *Analytical Methods 6*, 9 (2014), 2812–2831.

[BT03] BATTY M., TAYLOR M. J.: Early processing of the six basic facial emotional expressions. *Cognitive Brain Research 17*, 3 (2003), 613–620.

[CAE07] COHN J. F., AMBADAR Z., EKMAN P.: Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment* (2007), 203–221.

[CB03] CHIBELUSHI C. C., BOUREL F.: Facial expression recognition: A brief tutorial overview. *CVonline: On-Line Compendium of Computer Vision 9* (2003).

[CBJ00] CHEN H. F., BELHUMEUR P. N., JACOBS D. W.: In search of illumination invariants. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on* (2000), vol. 1, IEEE, pp. 254–261.

[CCG09] CAICEDO J. C., CRUZ A., GONZALEZ F. A.: Histopathology image classification using bag of features and kernel functions.

In *Conference on Artificial Intelligence in Medicine in Europe* (2009), Springer, pp. 126–135.

[CEW06] CHEN W., ER M. J., WU S.: Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 36*, 2 (2006), 458–466.

[CHS*93] CHIU K., HERF M., SHIRLEY P., SWAMY S., WANG C., ZIMMERMAN K., ET AL.: Spatially nonuniform scaling functions for high contrast images. In *Graphics Interface* (1993), CANADIAN INFORMATION PROCESSING SOCIETY, pp. 245–245.

[CMK*06] CARIDAKIS G., MALATESTA L., KESSOUS L., AMIR N., RAOUZAIOU A., KARPOUZIS K.: Modeling naturalistic affective states via facial and vocal expressions recognition. In *Proceedings of the 8th international conference on Multimodal interfaces* (2006), ACM, pp. 146–154.

[Coh07] COHN J. F.: Foundations of human computing: Facial expression and emotion. In *Artifical Intelligence for Human Computing*. Springer, 2007, pp. 1–16.

[CST00] CRISTIANINI N., SHAWE-TAYLOR J.: *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[CVBM02] CHAPELLE O., VAPNIK V., BOUSQUET O., MUKHERJEE S.: Choosing multiple parameters for support vector machines. *Machine learning 46*, 1-3 (2002), 131–159.

[CVL07] CHAN A. B., VASCONCELOS N., LANCKRIET G. R.: Direct convex relaxations of sparse svm. In *Proceedings of the 24th international conference on Machine learning* (2007), ACM, pp. 145–153.

[CYCC11] CHEN L.-H., YANG Y.-H., CHEN C.-S., CHENG M.-Y.: Illumination invariant feature extraction based on natural images statisticsâĂŤtaking face images as an example. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (2011), IEEE, pp. 681–688.

[CYZ*06] CHEN T., YIN W., ZHOU X. S., COMANICIU D., HUANG T. S.: Total variation models for variable lighting face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 28*, 9 (2006), 1519–1524.

[Dar72] DARWIN C.: 1965. the expression of the emotions in man and animals. *London, UK: John Marry* (1872).

[DBBB03] DRAPER B. A., BAEK K., BARTLETT M. S., BEVERIDGE J. R.: Recognizing faces with pca and ica. *Computer vision and image understanding 91*, 1 (2003), 115–137.

[DBDQ10] DUAN J., BRESSAN M., DANCE C., QIU G.: Tone-mapping high dynamic range images by novel histogram adjustment. *Pattern Recognition 43*, 5 (2010), 1847–1862.

[DBRS*15a] DEBATTISTA K., BASHFORD-ROGERS T., SELMANOVIĆ E., MUKHERJEE R., CHALMERS A.: Optimal exposure compression for high dynamic range content. *The Visual Computer 31*, 6-8 (2015), 1089–1099.

[DBRS*15b] DEBATTISTA K., BASHFORD-ROGERS T., SELMANOVIĆ E., MUKHERJEE R., CHALMERS A.: Optimal exposure compression

for high dynamic range content. *The Visual Computer 31*, 6-8 (2015), 1089–1099.

[DC05] DRBOHLAV O., CHANTLER M. J.: Two-image comparison under different illumination conditions. In *BMVC* (2005).

[DC10] DOUGLAS-COWIE E.: Mpeg-4 facial animation: The standard, implementation and applications.

[DD02] DURAND F., DORSEY J.: Fast bilateral filtering for the display of high-dynamic-range images. In *ACM transactions on graphics (TOG)* (2002), vol. 21, ACM, pp. 257–266.

[DDS*09] DENG J., DONG W., SOCHER R., LI L.-J., LI K., FEI-FEI L.: Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (2009), IEEE, pp. 248–255.

[Det15] DETTMERS T.: Understanding convolution in deep learning, 2015.

[DGLG11a] DHALL A., GOECKE R., LUCEY S., GEDEON T.: Acted facial expressions in the wild database. *Australian National University, Canberra, Australia, Technical Report TR-CS-11 2* (2011).

[DGLG11b] DHALL A., GOECKE R., LUCEY S., GEDEON T.: Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on* (2011), IEEE, pp. 2106–2112.

[DHS12] DUDA R. O., HART P. E., STORK D. G.: *Pattern classification*. John Wiley & Sons, 2012.

[DlTC11] DE LA TORRE F., COHN J. F.: Facial expression analysis. In *Visual analysis of humans*. Springer, 2011, pp. 377–409.

[DM08] DEBEVEC P. E., MALIK J.: Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes* (2008), ACM, p. 31.

[DMAC03] DRAGO F., MYSZKOWSKI K., ANNEN T., CHIBA N.: Adaptive logarithmic mapping for displaying high contrast scenes. In *Computer Graphics Forum* (2003), vol. 22, Wiley Online Library, pp. 419–426.

[DMMC03] DRAGO F., MARTENS W. L., MYSZKOWSKI K., CHIBA N.: Design of a tone mapping operator for high-dynamic range images based upon psychophysical evaluation and preference mapping. In *Electronic Imaging 2003* (2003), International Society for Optics and Photonics, pp. 321–331.

[EC97] ETEMAD K., CHELLAPPA R.: Discriminant analysis for recognition of human face images. *JOSA A 14*, 8 (1997), 1724–1733.

[EF71] EKMAN P., FRIESEN W. V.: Constants across cultures in the face and emotion. *Journal of personality and social psychology 17*, 2 (1971), 124.

[EF78] EKMAN P., FRIESEN W. V.: Facial action coding system: A technique for the measurement of facial action. *Manual for the Facial Action Coding System* (1978).

[EF03] EKMAN P., FRIESEN W. V.: *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

[EFE13] EKMAN P., FRIESEN W. V., ELLSWORTH P.: *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013.

[Ekm92] EKMAN P.: An argument for basic emotions. *Cognition & emotion 6*, 3-4 (1992), 169–200.

[EWMU13] EILERTSEN G., WANAT R., MANTIUK R. K., UNGER J.: Evaluation of tone mapping operators for hdr-video. In *Computer Graphics Forum* (2013), vol. 32, Wiley Online Library, pp. 275–284.

[FB15] FAN W., BOUGUILA N.: Face detection and facial expression recognition using simultaneous clustering and feature selection via an expectation propagation statistical learning framework. *Multimedia Tools and Applications 74*, 12 (2015), 4303–4327.

[FCY16] FOTIOS S., CASTLETON H., YANG B.: Does expression choice affect the analysis of light spectrum and facial emotion recognition? *Lighting Research and Technology* (2016), 1477153516651923.

[FDG*13] FANELLI G., DANTONE M., GALL J., FOSSATI A., VAN GOOL L.: Random forests for real time 3d face analysis. *International Journal of Computer Vision 101*, 3 (2013), 437–458.

[FE78] FRIESEN E., EKMAN P.: Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* (1978).

[FFFP07] FEI-FEI L., FERGUS R., PERONA P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding 106*, 1 (2007), 59–70.

[FFP05a] FEI-FEI L., PERONA P.: A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (2005), vol. 2, IEEE, pp. 524–531.

[FFP05b] FEI-FEI L., PERONA P.: A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (2005), vol. 2, IEEE, pp. 524–531.

[FKH] FLORIAN KAINZ ROD BOGART P. S., HILLMAN P.: *Technical Introduction to OpenEXR*. Industrial Light & Magic, Weta Digital.

[FL03] FASEL B., LUETTIN J.: Automatic facial expression analysis: a survey. *Pattern recognition 36*, 1 (2003), 259–275.

[FN09a] FRANCO A., NANNI L.: Fusion of classifiers for illumination robust face recognition. *Expert Systems with Applications 36*, 5 (2009), 8946–8954.

[FN09b] FRANCO A., NANNI L.: Fusion of classifiers for illumination robust face recognition. *Expert Systems with Applications 36*, 5 (2009), 8946–8954.

[FQ14] FARAJI M. R., QI X.: Face recognition under varying illumination with logarithmic fractal analysis. *Signal Processing Letters, IEEE 21*, 12 (2014), 1457–1461.

[FSYG10] FAN X., SUN Y., YIN B., GUO X.: Gabor-based dynamic representation for human fatigue monitoring in facial image sequences. *Pattern Recognition Letters 31*, 3 (2010), 234–243.

[GAKD00] GREENBERG S., ALADJEM M., KOGAN D., DIMITROV I.: Fingerprint image enhancement using filtering techniques. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on* (2000), vol. 3, IEEE, pp. 322–325.

[GB03] GROSS R., BRAJOVIC V.: An image preprocessing algorithm for illumination invariant face recognition. In *Audio-and Video-Based Biometric Person Authentication* (2003), Springer, pp. 10–18.

[Gil09] GILAT A.: *MATLAB: An introduction with Applications*. John Wiley & Sons, 2009.

[GJ10] GOPALAN R., JACOBS D.: Comparing and combining lighting insensitive approaches for face recognition. *Computer Vision and Image Understanding 114*, 1 (2010), 135–145.

[GNV11] GOEL T., NEHRA V., VISHWAKARMA V. P.: Comparative analysis of various illumination normalization techniques for face recognition. *International Journal of Computer Applications 28*, 9 (2011).

[Gol14] GOLLWITZER P. M.: Weakness of the will: Is a quick fix possible? *Motivation and Emotion 38*, 3 (2014), 305–322.

[GW08] GONZALEZ R. C., WOODS R. E.: Digital image processing. *Nueva Jersey* (2008).

[HL02] HSU C.-W., LIN C.-J.: A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on 13*, 2 (2002), 415–425.

[HS07] HONG J.-W., SONG K.-T.: Facial expression recognition under illumination variation. In *Advanced Robotics and Its Social Impacts, 2007. ARSO 2007. IEEE Workshop on* (2007), IEEE, pp. 1–6.

[HYAB13] HITAM M. S., YUSSOF W. N. J. H. W., AWALLUDIN E. A., BACHOK Z.: Mixture contrast limited adaptive histogram equalization for underwater image enhancement. In *Computer Applications*

Technology (ICCAT), 2013 International Conference on (2013), IEEE, pp. 1–5.

[HYH*05] HE X., YAN S., HU Y., NIYOGI P., ZHANG H.-J.: Face recognition using laplacianfaces. *IEEE transactions on pattern analysis and machine intelligence 27*, 3 (2005), 328–340.

[HYHH11] HU W.-C., YANG C.-Y., HUANG D.-Y., HUANG C.-H.: Feature-based face detection against skin-color like backgrounds with varying illumination. *Journal of Information Hiding and Multimedia Signal Processing 2*, 2 (2011), 123–132.

[IDC16] IGE E. O., DEBATTISTA K., CHALMERS A.: Towards hdr based facial expression recognition under complex lighting. In *Proceedings of the 33rd Computer Graphics International* (2016), ACM, pp. 49–52.

[IDMC16] IGE E. O., DEBATTISTA K., MUKHAJIE R., CHALMERS A.: Exploring face recognition under complex lighting conditions with hdr imaging. In *Proceedings of the Computer Graphics and Visual Computing* (2016), CGVC.

[IW79] IZARD C. E., ŒŢŒŹŒĄ ., WEISS M.: *Maximally discriminative facial movement coding system*. University of Delaware, instructional resources Center, 1979.

[JL10] JOHN LU Z.: The elements of statistical learning: data mining, inference, and prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 173*, 3 (2010), 693–694.

[JRW97] JOBSON D. J., RAHMAN Z.-U., WOODELL G. A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. *Image Processing, IEEE Transactions on 6*, 7 (1997), 965–976.

[JVP11] JIANG B., VALSTAR M. F., PANTIC M.: Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* (2011), IEEE, pp. 314–321.

[JWYh09] JIAN-WEI M., YU-HUA F.: Face segmentation algorithm based on asm. In *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on* (2009), vol. 4, IEEE, pp. 495–499.

[KA] KACHOURI R., AKIL M.: Conference 9400: Real-time image and video processing 2015. *TECHNICAL SUMMARIES*, 161.

[Kir09] KIRKEBØEN G.: Decision behaviour-improving expert judgement. In *Making essential choices with scant information*. Springer, 2009, pp. 169–194.

[KMM10] KALAL Z., MIKOLAJCZYK K., MATAS J.: Forward-backward error: Automatic detection of tracking failures. In *Pattern recognition (ICPR), 2010 20th international conference on* (2010), IEEE, pp. 2756–2759.

[KPB*13] KAHOU S. E., PAL C., BOUTHILLIER X., FROUMENTY P., GÜLÇEHRE Ç., MEMISEVIC R., VINCENT P., COURVILLE A., BENGIO Y., FERRARI R. C., ET AL.: Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (2013), ACM, pp. 543–550.

[KSMEK13] KODRA E., SENECHAL T., MCDUFF D., EL KALIOUBY R.: From dials to facial coding: Automated detection of spontaneous facial expressions for media research. In *Automatic Face and*

Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on (2013), IEEE, pp. 1–6.

[KSZ*14] KEINERT J., SCHÖBERL M., ZIEGLER M., ZILLY F., FOESSEL S.: High-dynamic range video cameras based on single shot non-regular sampling. *SMPTE Motion Imaging Journal 123*, 8 (2014), 49–54.

[KYJF04] KUANG J., YAMAGUCHI H., JOHNSON G. M., FAIRCHILD M. D.: Testing hdr image rendering algorithms. In *Color and Imaging Conference* (2004), vol. 2004, Society for Imaging Science and Technology, pp. 315–320.

[LAKG98] LYONS M., AKAMATSU S., KAMACHI M., GYOBA J.: Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on* (1998), IEEE, pp. 200–205.

[LHK05] LEE K.-C., HO J., KRIEGMAN D. J.: Acquiring linear subspaces for face recognition under variable lighting. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 27*, 5 (2005), 684–698.

[LLSC13] LIU M., LI S., SHAN S., CHEN X.: Au-aware deep networks for facial expression recognition. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on* (2013), IEEE, pp. 1–6.

[LLSZ15] LI W., LI M., SU Z., ZHU Z.: A deep-learning approach to facial expression recognition with candid images. In *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on* (2015), IEEE, pp. 279–282.

[LM01] LEUNG T., MALIK J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. *International journal of computer vision 43*, 1 (2001), 29–44.

[LMMHM01] LENNON M., MERCIER G., MOUCHOT M., HUBERT-MOY L.: Independent component analysis as a tool for the dimensionality reduction and the representation of hyperspectral images. In *Geoscience and Remote Sensing Symposium, 2001. IGARSS'01. IEEE 2001 International* (2001), vol. 6, IEEE, pp. 2893–2895.

[LPR14] LEE S. H., PLATANIOTIS K. N. K., RO Y. M.: Intra-class variation reduction using training expression images for sparse representation based facial expression recognition. *IEEE Transactions on Affective Computing 5*, 3 (2014), 340–351.

[LWY*13] LI Q., WANG H. J., YOU J., LI Z. M., LI J. X.: Enlarge the training set based on inter-class relationship for face recognition from one image per person. *PloS one 8*, 7 (2013), e68539.

[LZY12] LU Y., ZHOU J., YU S.: A survey of face detection, extraction and recognition. *Computing and informatics 22*, 2 (2012), 163–195.

[MBDC15] MELO M., BESSA M., DEBATTISTA K., CHALMERS A.: Evaluation of tone-mapping operators for hdr video under different ambient luminance levels. *Computer Graphics Forum 34*, 8 (2015), 38–49.

[MCM16] MOLLAHOSSEINI A., CHAN D., MAHOOR M. H.: Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2016), IEEE, pp. 1–10.

[MDK08a] MANTIUK R., DALY S., KEROFSKY L.: Display adaptive tone mapping. *ACM Transactions on Graphics (TOG) 27*, 3 (2008), 68.

[MDK08b] MANTIUK R., DALY S., KEROFSKY L.: Display adaptive tone mapping. In *ACM SIGGRAPH 2008 Papers* (New York, NY, USA, 2008), SIGGRAPH '08, ACM, pp. 68:1–68:10.

[Meh08] MEHRABIAN A.: Communication without words. *Communication Theory,* (2008), 193–200.

[MEK03] MICHEL P., EL KALIOUBY R.: Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th international conference on Multimodal interfaces* (2003), ACM, pp. 258–264.

[mel]

[MG] MOHAMMED S. N., GEORGE L. E.: Illumination-invariant facial components extraction using adaptive contrast enhancement methods.

[MJW*13] MIENALTOWSKI A., JOHNSON E. R., WITTMAN R., WILSON A.-T., STURYCZ C., NORMAN J. F.: The visual discrimination of negative facial expressions by younger and older adults. *Vision research 81* (2013), 12–17.

[MM] MA R., MOHAMED A.: Image processing pipeline for facial expression recognition under variable lighting.

[MMK08] MYSZKOWSKI K., MANTIUK R., KRAWCZYK G.: High dynamic range video. *Synthesis Lectures on Computer Graphics and Animation 1*, 1 (2008), 1–158.

[MMMK03] MATSUGU M., MORI K., MITARI Y., KANEDA Y.: Subject independent facial expression recognition with robust face

detection using a convolutional neural network. *Neural Networks 16*, 5 (2003), 555–559.

[Mür15] MÜRI R. M.: Cortical control of facial expression. *Journal of Comparative Neurology* (2015).

[NMP98] NEUMANN L., MATKOVIĆ K., PURGATHOFER W.: Automatic exposure in computer graphics based on the minimum information loss principle. In *Computer Graphics International, 1998. Proceedings* (1998), IEEE, pp. 666–677.

[NNVW15] NG H.-W., NGUYEN V. D., VONIKAKIS V., WINKLER S.: Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (2015), ACM, pp. 443–449.

[NPDlT08] NGUYEN M. H., PEREZ J., DE LA TORRE F.: Facial feature detection with optimal pixel reduction svm. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on* (2008), IEEE, pp. 1–6.

[OPAHC*14a] OUSSALAH M., PROFESSOR ALI HESSAMI D., CHERMAK L., AOUF N., RICHARDSON M.: Hdr imaging for feature tracking in challenging visibility scenes. *Kybernetes 43*, 8 (2014), 1129–1149.

[OPAHC*14b] OUSSALAH M., PROFESSOR ALI HESSAMI D., CHERMAK L., AOUF N., RICHARDSON M.: Hdr imaging for feature tracking in challenging visibility scenes. *Kybernetes 43*, 8 (2014), 1129–1149.

[PCB01] PHUNG S. L., CHAI D., BOUZERDOUM A.: Skin colour based face detection. In *Intelligent Information Systems Conference, the Seventh Australian and New Zealand 2001* (2001), IEEE, pp. 171–176.

[PFW*11] PONCE J., FORSYTH D., WILLOW E.-P., ANTIPOLIS-MÉDITERRANÉE S., DÂĂŹACTIVITÉ RAWEB R., INRIA L., ALUMNI I.: Computer vision: a modern approach. *Computer 16*, 11 (2011).

[Plu01] PLUTCHIK R.: The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist 89*, 4 (2001), 344–350.

[PMPP14] PEREIRA M., MORENO J.-C., PROENÇA H., PINHEIRO A. M.: Automatic face recognition in hdr imaging. In *SPIE Photonics Europe* (2014), International Society for Optics and Photonics, pp. 913804–913804.

[PP06] PANTIC M., PATRAS I.: Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 36*, 2 (2006), 433–449.

[PSO*07] PHILLIPS P. J., SCRUGGS W. T., OÂĂŹTOOLE A. J., FLYNN P. J., BOWYER K. W., SCHOTT C. L., SHARPE M.: Frvt 2006 and ice 2006 large-scale results. *National Institute of Standards and Technology, NISTIR 7408* (2007), 1.

[PYL08] PANG Y., YUAN Y., LI X.: Gabor-based region covariance matrices for face recognition. *IEEE Transactions on Circuits and Systems for Video Technology 18*, 7 (2008), 989–993.

[RAI12] ROUHI R., AMIRI M., IRANNEJAD B.: A review on feature extraction techniques in face recognition. *Signal & Image Processing: An International Journal (SIPIJ)* (2012).

[RB09] RODRIGO M. M. T., BAKER R. S.: Coarse-grained detection of student frustration in an introductory programming course. In *Proceedings of the fifth international workshop on Computing education research workshop* (2009), ACM, pp. 75–80.

[RBFD03] RUSSELL J. A., BACHOROWSKI J.-A., FERNÁNDEZ-DOLS J.-M.: Facial and vocal expressions of emotion. *Annual review of psychology 54*, 1 (2003), 329–349.

[RdSQ08] RUIZ-DEL SOLAR J., QUINTEROS J.: Illumination compensation and normalization in eigenspace-based face recognition: A comparative study of different pre-processing approaches. *Pattern Recognition Letters 29*, 14 (2008), 1966–1979.

[Rei02] REINHARD E.: Parameter estimation for photographic tone reproduction. *Journal of graphics tools 7*, 1 (2002), 45–51.

[RJW04] RAHMAN Z.-U., JOBSON D. J., WOODELL G. A.: Retinex processing for automatic image enhancement. *Journal of Electronic Imaging 13*, 1 (2004), 100–110.

[RPP12] RUDOVIC O., PAVLOVIC V., PANTIC M.: Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (2012), IEEE, pp. 2634–2641.

[RR] REDDY K. S., REDDY K. R. L.: Enlargement of image based upon interpolation techniques.

[RWHJ13] ROBINSON M. D., WATKINS E. R., HARMON-JONES E.: *Handbook of cognition and emotion*. Guilford Press, 2013.

[SAK*15] SIDDIQI M. H., ALI R., KHAN A. M., PARK Y.-T., LEE S.: Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields. *IEEE Transactions on Image Processing 24*, 4 (2015), 1386–1398.

[SB10] SHAN C., BRASPENNING R.: Recognizing facial expressions automatically from video. In *Handbook of ambient intelligence and smart environments*. Springer, 2010, pp. 479–509.

[SBB02] SIM T., BAKER S., BSAT M.: The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on* (2002), IEEE, pp. 46–51.

[Sch95] SCHLICK C.: Quantization techniques for visualization of high dynamic range pictures. In *Photorealistic Rendering Techniques*. Springer, 1995, pp. 7–20.

[SGCZ03] SHAN S., GAO W., CAO B., ZHAO D.: Illumination normalization for robust face recognition against varying lighting conditions. In *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on* (2003), IEEE, pp. 157–164.

[SGDM11] STRATOU G., GHOSH A., DEBEVEC P., MORENCY L.-P.: Effect of illumination on automatic expression recognition: a novel 3d relightable facial database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* (2011), IEEE, pp. 611–618.

[SGM09] SHAN C., GONG S., MCOWAN P. W.: Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing 27*, 6 (2009), 803–816.

[Sha97] SHASHUA A.: On photometric issues in 3d visual recognition from a single 2d image. *International Journal of Computer Vision 21*, 1-2 (1997), 99–122.

[SHS*04] SEETZEN H., HEIDRICH W., STUERZLINGER W., WARD G., WHITEHEAD L., TRENTACOSTE M., GHOSH A., VOROZCOVS A.: High dynamic range display systems. *ACM Transactions on Graphics (TOG) 23*, 3 (2004), 760–768.

[SI92] SAMAL A., IYENGAR P. A.: Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern recognition 25*, 1 (1992), 65–77.

[SKM04] SHORT J., KITTLER J., MESSER K.: A comparison of photometric normalisation algorithms for face verification. In *null* (2004), IEEE, p. 254.

[SL09] SOKOLOVA M., LAPALME G.: A systematic analysis of performance measures for classification tasks. *Information Processing & Management 45*, 4 (2009), 427–437.

[SM11] SCHMITT D., MCCOY N.: Object classification and localization using surf descriptors. *CS 229* (2011), 1–5.

[SMCD03] STARCK J.-L., MURTAGH F., CANDES E. J., DONOHO D. L.: Gray and color image contrast enhancement by the curvelet transform. *IEEE Transactions on image processing 12*, 6 (2003), 706–717.

[SSM12] SUMATHI C., SANTHANAM T., MAHADEVI M.: Automatic facial expression analysis a survey. *International Journal of Computer Science and Engineering Survey 3*, 6 (2012), 47.

[STL*10] SONG M., TAO D., LIU Z., LI X., ZHOU M.: Image ratio features for facial expression recognition application. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 40*, 3 (2010), 779–788.

[TKC01] TIAN Y.-I., KANADE T., COHN J. F.: Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence 23*, 2 (2001), 97–115.

[TKC05] TIAN Y.-L., KANADE T., COHN J. F.: Facial expression analysis. In *Handbook of face recognition*. Springer, 2005, pp. 247–275.

[TT10] TAN X., TRIGGS B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Image Processing, IEEE Transactions on 19*, 6 (2010), 1635–1650.

[UMM*10] URBANO C., MAGALHÃČES L., MOURA J., BESSA M., MARCOS A., CHALMERS A.: Tone mapping operators on small screen devices: An evaluation study. *Computer Graphics Forum 29*, 8 (2010), 2469–2478.

[VDJ10] VAVILIN A., DEB K., JO K.-H.: Fast hdr image generation technique based on exposure blending. In *Trends in Applied Intelligent Systems*. Springer, 2010, pp. 379–388.

[VGV*16] VERMA N. K., GOYAL A., VARDHAN A. H., SEVAKULA R. K., SALOUR A.: Object matching using speeded up robust features. In *Intelligent and Evolutionary Systems*. Springer, 2016, pp. 415–427.

[VJ01] VIOLA P., JONES M.: Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* (2001), vol. 1, IEEE, pp. I–511.

[VL15] VEDALDI A., LENC K.: Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia* (2015), ACM, pp. 689–692.

[VP12] VALSTAR M. F., PANTIC M.: Fully automatic recognition of the temporal phases of facial actions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 42*, 1 (2012), 28–43.

[VPG09] VISHWAKARMA V. P., PANDEY S., GUPTA M.: Adaptive histogram equalization and logarithm transform with rescaled low frequency dct coefficients for illumination normalization. *International Journal of Recent Trends in Engineering 1*, 1 (2009).

[WCL11] WANG B., CHANG X., LIU C.: Skin detection and segmentation of human face in color images. *International Journal of Intelligent Engineering and Systems 4*, 1 (2011), 10–17.

[WLF*09] WHITEHILL J., LITTLEWORT G., FASEL I., BARTLETT M., MOVELLAN J.: Toward practical smile detection. *IEEE transactions on pattern analysis and machine intelligence 31*, 11 (2009), 2106–2111.

[WLH03] WEN Z., LIU Z., HUANG T. S.: Face relighting with radiance environment maps. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on* (2003), vol. 2, IEEE, pp. II–158.

[WLH*07] WANG Y., LIU Z., HUA G., WEN Z., ZHANG Z., SAMARAS D.: Face re-lighting from a single image under harsh lighting conditions. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (2007), IEEE, pp. 1–8.

[WLL*10] WANG S., LIU Z., LV S., LV Y., WU G., PENG P., CHEN F., WANG X.: A natural visible and infrared facial expression database for

expression recognition and emotion inference. *IEEE Transactions on Multimedia 12*, 7 (2010), 682–691.

[WMC*00] WESTON J., MUKHERJEE S., CHAPELLE O., PONTIL M., POGGIO T., VAPNIK V.: Feature selection for svms.

[WPB11] WILLIS M. L., PALERMO R., BURKE D.: Judging approachability on the face of it: The influence of face and body expressions on the perception of approachability. *Emotion 11*, 3 (2011), 514.

[YCBL14] YOSINSKI J., CLUNE J., BENGIO Y., LIPSON H.: How transferable are features in deep neural networks? In *Advances in neural information processing systems* (2014), pp. 3320–3328.

[YF14] YANG B., FOTIOS S.: Lighting and recognition of emotion conveyed by facial expressions. *Lighting Research and Technology* (2014), 1477153514547753.

[ZBMM06] ZHANG H., BERG A. C., MAIRE M., MALIK J.: Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (2006), vol. 2, IEEE, pp. 2126–2136.

[Zha10] ZHANG Y.-J.: *Advances in Face Image Analysis: Techniques and Technologies: Techniques and Technologies*. IGI Global, 2010.

[ZHR*07] ZENG Z., HU Y., ROISMAN G. I., WEN Z., FU Y., HUANG T. S.: Audio-visual spontaneous emotion recognition. In *Artifical Intelligence for Human Computing*. Springer, 2007, pp. 72–90.

[ZLSA98] ZHANG Z., LYONS M., SCHUSTER M., AKAMATSU S.: Comparison between geometry-based and gabor-wavelets-based facial

expression recognition using multi-layer perceptron. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on* (1998), IEEE, pp. 454–459.

[ZPRH09] ZENG Z., PANTIC M., ROISMAN G. I., HUANG T. S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence 31*, 1 (2009), 39–58.

[ZS03] ZHANG L., SAMARAS D.: Face recognition under variable lighting using harmonic image exemplars. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on* (2003), vol. 1, IEEE, pp. I–19.

[ZSCG07] ZHANG B., SHAN S., CHEN X., GAO W.: Histogram of gabor phase patterns (hgpp): a novel object representation approach for face recognition. *Image Processing, IEEE Transactions on 16*, 1 (2007), 57–68.

[ZTC14a] ZHANG L., TJONDRONEGORO D., CHANDRAN V.: Facial expression recognition experiments with data from television broadcasts and the world wide web. *Image and Vision Computing 32*, 2 (2014), 107–119.

[ZTC14b] ZHANG L., TJONDRONEGORO D., CHANDRAN V.: Facial expression recognition experiments with data from television broadcasts and the world wide web. *Image and Vision Computing 32*, 2 (2014), 107–119.

[ZZL*15] ZHAO X., ZOU J., LI H., DELLANDRÉA E., KAKADIARIS I. A., CHEN L.: Automatic 2.5-d facial landmarking and emotion annotation for social interaction assistance.